Murray Shanahan

THE TECHNOLOGICAL SINGULARITY

技术奇点

当机器拥有人性，我们将面对怎样的世界？

揭开人工智能的神秘面纱，激发重塑人类社会的力量

[英] 默里·沙纳汉 著

霍斯亮 译



中信出版集团 · CHINA CITIC PRESS

版权信息

书名:技术奇点

作者:默里·沙纳汉

译者:霍斯亮

ISBN:9787508659701

中信出版集团制作发行

题记

本书的内容对一些读者来说可能很怪诞，但是对于作者来说，这都是真实而迫切的，绝不仅仅是科幻小说的内容，应该被认真对待。

——I·J·古德（I. J. Good），《关于第一台超智能机器的推测》（Speculations Concerning of the First Ultraintelligent Machine），1965年

人工智能最大的问题在于没有动机。它们毕竟不是人，明白吗？

——威廉·吉布森（William Gibson），《神经漫游者》（Neuromancer），1984年

前言

近年来，随着科学技术的迅猛发展，“奇点”这个概念已经从科幻小说逐渐转移到了严肃的学术讨论中。人们认为当前人类历史的发展阶段距离奇点已越来越近。物理学中的奇点指的是在时间或空间的某一点上，出现了类似黑洞或者宇宙大爆炸的情况，数学已经不再适用，人类也无法理解。人类历史中出现的奇点指的是由于技术的迅速发展，人类社会中的一切都出现了改变，生活在今天的我们将无法理解。注今天被我们当作理所当然的一切（经济、政府、法律、国家）会改变形态，人类最根本的价值观（生命的尊严、对幸福的追求、选择的自由）也会被淘汰。我们对人类之所以为人的理解（人性、生存、意识、社会秩序）提出疑问，不是通过漫无边际的哲学思考，而是由真实存在的环境所推动。

什么样的技术进步会带来这种剧变？本书将要验证的假说是：人工智能和神经技术这两个领域的技术进步会加速奇点的到来。我们已经破解了生命的秘密，能够理解基因和DNA的工作原理。生物技术带来的改变已经非常巨大，但是和人工制造思维的能力相比，就差多了。

从某种重要的意义上来说，今天的智能是固定的，这影响了技术进步的广度和速度。几千年来，人们掌握的知识已经非常丰富，而且随着文字、印刷术和互联网的出现，传播知识的能力也增强了。但是，创造知识的器官（有智慧的人的大脑）在过去的几千年中变化不大，认知能力的秘密还没有被我们揭开。

如果人工智能和神经科学能充分发挥潜力，这种情况将会改变。如果不仅是智能创造技术，技术反过来也可以创造智能，那么就会产

生一个循环。这个循环的结果是难以预料的，可能带来爆炸性的改变。作为原来的创造者，如果智能也可以被创造出来，那么它就可以提升自己。根据奇点理论，不久以后，这个循环中就不再需要人类了，人工智能或进行过认知改造的生物智能将占据主导地位，人类已经跟不上了。

我们需要认真对待奇点理论的假说吗？又或者它只适合出现在科幻小说里？雷·库兹韦尔认为我们应该认真对待这一假说，并提出了“加速回报定律”（The Law of Accelerated Return）。如果一项科技符合加速回报定律，那么这项科技越先进，它进步的速度就越快，在一段时间以后就会实现指数级进步。

一个很好的例子就是“摩尔定律”（Moore’s Law），即一块芯片能容纳的晶体管数量每隔18个月左右就会翻番。令人惊讶的是，半导体行业几十年来的发展趋势一直遵循摩尔定律。信息技术发展的其他指标，例如CPU（中央处理器）的时钟频率和网络带宽，也一直在以指数级速度增长。然而，信息技术不是唯一高速发展的领域。在医学领域，DNA（脱氧核糖核酸）测序的成本快速下降，测序的速度则呈指数级上升，大脑扫描的分辨率也在以指数级速度上升。

作为原来的创造者，

如果智能也可以被创造出来，

那么它就可以提升自己。

从历史的角度来看，出现重大技术进步的时间间隔在缩短：从农业、印刷术、电力到电脑。除了技术的进步，进化过程中出现里程碑式突破的间隔也在不断缩短：从真核生物、脊椎动物、灵长类动物到智人。很多评论家认为人类发展的轨迹是一条复杂的曲线，一直延伸到远古时代。但是，我们只需要研究这条曲线中科技的部分，并推演科技曲线的未来趋势。我们发现，未来会出现一个转折点，人类发明的技术会让自己落伍。

由于物理规律的制约，指数级技术发展最终会遇到瓶颈。指数级技术发展的停滞可能有各种经济、政治或者科学的原因。但是，我们先假设与人工智能以及神经技术最相关的发展将会保持势头，加快破解大脑秘密的速度，并且最终合成人工智能。我们不难相信，不管是人工智能还是人类智能，都会遵循加速回报定律，最终到达技术奇点。

有些作者预言，奇点的到来会在21世纪中叶。我们可以不把奇点只作为预言来研究，预言毕竟是比较虚无缥缈的。首先，奇点是一个很有意思的深刻的学术话题，无论它最后是否真的会出现或不知什么时候出现。其次，奇点出现的可能性虽然看起来很微小，但是却值得我们从务实的角度，在理性的基础上讨论它。即使未来学的理论有瑕疵，但只要未来学描述的场景存在可能性，就值得我们认真关注。如果奇点真的出现，对人类来讲，其后果将是颠覆性的。

这些颠覆性的后果都有哪些？如果人类到达奇点，世界会变成什么样子？我们对奇点的到来是应该恐惧，还是应该欢迎？人类在今天或未来应该做些什么来保证最好的结果？本书旨在回答这一系列非常宏大的问题。奇点这个概念会引发我们对一些更加原始的哲学问题的讨论：人性的本质是什么？人类最根本的价值观是什么？我们应该如何生存？人类为了迎接奇点需要放弃什么？奇点在给人类的生存带来机遇的同时，也带来挑战。

奇点的到来可能会威胁到人类的生存。这听起来像是夸大其词，但是今天的新兴技术已经超越了以往，是人类从来没有见过的。可能会有人研制经过基因改造的病毒，提高其传染性和抗药性——只有疯子才会做这种事，但是，人类有可能意外地制造出带来灾难的病毒。高级人工智能也有可能带来生存威胁，其原因相似，但更微妙。我们会在后文中研究这些问题。如果未来某个公司、政府、组织，甚至是个人有可能制造出能实现指数级自我提升、极度渴求资源的人工智能，这一可能性足以让我们提高警惕。

从乐观的角度来看，技术奇点带来了生存的机会，也就是哲学上的“存在”的意义。当我们有能力构建大脑时，就可以突破生物的限制，延长人类的寿命。人类的最终极限就是死亡，生物的身体是很脆弱的，会受到疾病、意外伤害和衰老的限制。人类的意识依赖其生物大脑，但是，未来人类将有能力修复任何对大脑的损害，并最终在其他基质上构建大脑，这样就可以突破意识的寿命极限，将其无限延长。

延长生命是“超级人类主义”的一个分支，但是我们为什么要满足于人类的大脑？如果我们能够重建大脑，为什么不通过重新设计，提升大脑的能力呢（人类的身体也可以通过重建提升，但是本书只讨论智能）？通过医药手段可以提高记忆力、学习能力和注意力，这是比较传统的提升方法。但是，既然人类能够从无到有重建大脑，就一定能对认知能力进行重组和提升。当我们具备了这种提升能力，又该做什么呢？有些人认为，这种能力至少能够应对超级智能带来的存在风险，赶上人工智能的发展脚步，不过在这个过程中人类的认知方式可能会发生根本改变。

技术奇点可能是一种生存机会，要理解这一概念，必须放弃人类的视角，采取一种更具宇宙性的视角。“人类社会和生活在人类社会中的人们的大脑已经到达了宇宙的极限”是一种人类中心主义的想

法。也许智能的复杂程度还有很大的提升空间。可能未来还会出现比我们更高级的意识形式。我们应该阻止还是迎接这一情况的到来？我们能理解这样的未来会是什么样吗？不管奇点是否已经临近，这些问题都值得思考。因为在思考这些问题的同时我们对自身有了新的理解，也理解了我们在大千世界中的位置。

---

1. “奇点”这个词是由约翰·冯·诺伊曼（John Von Neumann）首次提出的。雷·库兹韦尔（Ray Kurzweil）在他2005年的著作《奇点临近》（The Singularity is Near）中将“奇点”一词推广开来。这个词在当代有多种含义，库兹韦尔使用的含义与弗诺·文奇（Vernor Vinge）在《即将到来的技术奇点》（The Coming Technological Singularity）一文中使用的含义最接近。

THE TECHNOLOGICAL SINGULARITY

第一章

人工智能，机器与人类的博弈

当机器开始学习

艾伦·图灵（Alan Turing）是第二次世界大战期间的密码破译专家，计算机科学的先驱。1950年，他在《心灵》（Mind）杂志上发表了一篇文章，名为“计算机与智能”（Computing Machinery and Intelligence），这是第一篇真正对人工智能进行严肃学术讨论的文章。图灵预测道：“到2000年，人们在谈到机器也能思考时，会觉得是一件很正常的事。”他预言机器将能够通过“图灵测试”。

图灵测试是一种游戏，有两个“玩家”，一个是人类，一个是机器，“玩家”通过键盘和屏幕与人类“裁判”交流。“裁判”轮流与两个“玩家”对话，试图猜出哪个是人类，哪个是机器。机器的任务是让“裁判”相信它才是人类，而这绝对需要人类水平的智力。如果“裁判”无法分辨哪个是人类，那么机器就通过了测试。这篇文章写于1950年。当时，图灵预测未来所有普通的机器都能通过这种测试。在未来的世界里，“能思考的机器”在家庭和工作环境中都很普遍。

虽然图灵做出了这样的预测，但是实际上在2000年，具有人类智慧水平的机器没有出现，也没有迹象表明人类能在短期内实现这一目标。现在计算机的水平距离通过图灵测试还很遥远。但是，近年来人工智能开发取得了里程碑式的发展。1997年，IBM（国际商业机器公司）开发的电脑“深蓝”击败了世界国际象棋冠军加里·卡斯帕罗夫（Gary Kasparov）。据说，卡斯帕罗夫曾经表示，“深蓝”和他以往击败的电脑迥然不同。以往的电脑对手令人感觉平庸机械，而在和“深蓝”对战的时候，卡斯帕罗夫感觉在棋盘另一端的是某种“外星智慧生物”。

我们应当仔细思考这一里程碑事件的意义。半个世纪以前，这一事件被认为是人工智能领域最大的成就——机器战胜了人类。当然，汽车的速度比人类短跑冠军更快，起重机的力气比举重冠军更大，但是智力的优越是人类与其他动物的根本区别，国际象棋正是纯粹智力上的比赛。

现在，人类在国际象棋领域输给了电脑。但是，与图灵的时代相比，我们与人类水平人工智能的距离似乎并没有缩短多少。这是为什么呢？因为“深蓝”是只有一种本领的专才，它只会下象棋，而人类则正好相反。比如，我正在咖啡馆里用电脑写文章，一位女白领经过我的窗前。她在一天中要做各种各样的事情：做午餐便当、检查孩子的作业、开车上班、写电子邮件、修理打印机等。如果我们仔细研究这一系列活动，会发现它们都需要人类具备感知运动的能力。例如做午餐便当，就需要从各处拿厨具、食材，打开包装、做饭、把饭装进饭盒等等。

简言之，人类是通才，是什么都会做的万事通。人类国际象棋选手除了会下棋之外，还会做各种事情。此外，人类还有适应能力。人不是天生就会修理打印机，而是后天学会的。那位女白领如果出生在另一个世纪的另一种文化的国家，可能会学习一系列不同的技能。如果她不幸丢了工作，也必须重新接受培训，学习新的技能。人工智能研究在很多专业领域（例如国际象棋）取得了成功，但是却没能发明一种万事通机器，像人类一样什么都会，还能适应环境。我们如何发明强人工智能[注]？只有弄清这个问题，才能讨论超级人工智能。

生物智能的主要特点是有实体。和“深蓝”相反，人类实际上是一种动物，有着血肉之躯，大脑是身体的一部分。动物的大脑经过进化获得了控制身体的能力，使身体能够把基因传递下去。肌肉让身体能够运动，感觉让身体能根据周围的环境调整动作，完成大脑的指令。大脑坐镇于感觉运动（sensorimotor）的环路之中，根据得到的

信息，控制着动物的行动。不管人类大脑的成就有多大，从根本上来说，也只是动物大脑的一种升级。人类的语言能力、推理能力和创造力都根植于感觉运动这一基础。

创造人工智能可以省去有机生物必需的很多活动，比如新陈代谢和繁殖，但是不能省去人工智能的实体。正是为了应对复杂多变的物理世界，才诞生了智能——毕竟这个世界充满多种多样的物体，有的是静止的，有的是活动的。从这个角度来讲，图灵测试是不完美的，因为图灵测试只涉及语言。判断人工智能的有效方法是看它如何与我们所处的环境互动。因此，要创造真正的人类水平人工智能，就必须制造机器人。但是，对于机器人这种实体，也有人反对。我们之后会讨论反对的观点，现在先假定这一判断是正确的。那么，我们面对的问题就是：如何让机器人拥有强人工智能？

强人工智能可能就是将很多感觉运动的专业能力结合在一起，但是现在我们还没有研发出足够的专业能力。那么，当人们赋予机器人足够多样的能力时，就能产生强人工智能了吗？即使我们忽略工程方面的困难，这一假设仍然难以让人信服。使用这一方法制造出的机器人可能看上去很像强人工智能，但是很难蒙混过关。这种多功能机器人一旦遇到能力范围之外的问题，马上就会束手无策。世界变化这么快，肯定会出现它能力之外的问题。

只有当人工智能拥有学习的能力时，或许才能跨越这一障碍。在不熟悉的环境中，人工智能应当能够学习新的技能。只有拥有了学习能力，才能获得和维持新技能。事实上，学习的能力，不管是什么形式的学习，都是建立智能的基础。但是，学习是非常耗费时间的，也是充满风险的。拥有普通智力的标志应当是在新挑战面前，不需要经过尝试或第三方的培训就能采用新的行为模式去适应环境。

---

1. 强人工智能（Aritificial General Intelligence，简称AGI）：具备一般智慧能力的人工智能，有知觉和意识。——译者注

常识与创造力，让机器了解世界

如何克服专业能力的缺陷，让机器拥有普通智力？最基本的要求或许就是拥有常识和创造力。常识在这里指的是能够理解日常生活中物理环境和社会环境的运作规律。例如，一个人绕着一座建筑物走一圈，会回到原点。常识就是这样的道理。又例如，一个人沿着一条路向前走，走到头折返回来，回程会看到刚才见过的建筑物，但是顺序相反。这类道理非常有用，因为应用的范围非常广泛，具有普遍性和永久性。

拥有常识意味着什么？拥有常识不代表理解事物运行的规律，也不代表能把规律用语言总结出来。常识主要是能指导行动。换句话说，如果缺乏常识，就会在行动上表现出来。

例如，我家屋后的大公鸡喜欢逃出鸡圈，飞过大门，不过它总是在外面玩一会儿就想回鸡圈了。其实它只需要按原路从大门飞回来就可以了，但是它似乎总是想不到这一点，只能在大门口焦急地踱来踱去。有些行为是可逆的——大公鸡似乎就没有这个常识。

动物的行为中虽然会有与上文类似的盲点，但也算拥有一定的常识。人类也拥有常识，而且常识还会扩展到社会层面。例如，人们通过语言交流自己对世界的理解。假设你早上到了公司以后，发现一群同事站在办公楼外面淋雨。你问其中一个人：“你们干什么呢？”如果她按照字面意思回答“我们在淋雨”，你可能会觉得很奇怪。如果她说“着火了”，虽然没有直接回答你的问题，但是她明白人们对话的目的是交换信息。这就是一种常识。

普通智力的另一项要求是创造力。这里说的创造力不是艺术家、作曲家或者数学家的创造力，而是普通人也有的创造力，儿童特别富于这种创造力。这种创造力指的是创造新的行为方式、发明新事物或者用新方法使用旧东西的能力。这种创造力可能完全是探索性的或者只是为了好玩，就好像儿童自己手舞足蹈，其实就是创造了一种新舞蹈。这种创造力也可能有着明确的目标，比如设计一个花园，或者想办法减少家庭支出。这些似乎都是非常普通的人类行为，但是却需要一个人打破原有的行为模式，建立新的行为模式，或者把原有的行为模式重新组合。

创造力和常识是相互补充的。一个人依靠创造力做出的新行为，需要通过关于日常世界的常识来预测其结果。一方面，只有想象力没有常识会让人像没头苍蝇一样乱撞，另一方面，只有常识没有想象力会让人僵化死板。但是，二者兼备的智能是非常强大的。在面对新挑战时，这种智能可以用想象力设计大量可能的行为，然后用常识来预测这些行为的效果。不管是人还是人工智能，在运动肌肉或者开动发动机之前，对自己的行为带来的结果是心里有数的。

2002年，亚历克斯·卡采尔尼克（Alex Kacelnik）领导的牛津大学动物认知研究人员发表了他们的研究成果：他们观察到了动物身上表现出的创新行为。他们以新喀里多尼亚乌鸦（一种特别聪明的乌鸦）为研究对象做了一项实验。他们把一小桶食物放在一个立着的深管子里。乌鸦在管子外面够不着桶或者桶的提手。在研究人员为乌鸦提供一段弯曲的铁丝后，乌鸦很快就学会把铁丝用作钩子把桶钩上来。随后研究人员为乌鸦提供了一段直的铁丝。一只名叫贝蒂的乌鸦在没有经过任何训练的情况下，把铁丝塞进设备的一个小孔里，将铁丝弯折成钩子的形状，然后把桶提了上来。

贝蒂的行为既富于创造力，又体现了常识。创造力让贝蒂产生了将原本无用的铁丝折弯的想法，同时，因为拥有常识，所以它了解铁

丝可以被折弯的属性。这两种认知的元素在动物身上产生了神奇的效果，在具有语言能力的人类身上效果将会更加明显。比如，一个男中学生用一句俏皮话去取笑他的同学，这既需要语言的创造力，又需要关于人类心理的基本常识（当然更重要的常识是不能这样去取笑老师）。这只是一个小例子。但是，所有人类的成就，从金字塔到登月，都是无数个这种发明创造的集合。人类水平人工智能只有同时具有常识和创造力，才能取得这种水平的成就。

实现人工智能，模仿还是创造

如果创造人工智能的要求这么明确——一点创造力和一点常识，为什么这个领域60年来的研究没有进步？既然这一领域没有取得什么成绩，我们还有可能实现人类水平人工智能吗？如果创造人类水平人工智能都这么困难，还有什么必要讨论超级智能呢？之前我们一直在讨论普通智力的行为模式，没有谈到实现智力的机制——不管是生物大脑的智力还是人造的智能。但是，要回答这些问题，我们就不能回避实现普通智力的机制。要讨论人工智能的未来，必须要提到使用什么样的机制来实现。从计算机科学的角度来讲就是：我们不仅要谈规范，还要谈如何实现。

在计算机科学方面，一个规范通常可以有很多种实现方法。这样就比较复杂了。软件公司只需要做出一款能满足需求的软件，而我们要考虑人工智能所有的可能性。此外，未来也许会出现一种革命性的新技术，以我们现在难以想象的方式实现人工智能。所以，我们只能挨个研究现在人工智能的主要流派，并试图得出一些结论。

要讨论人工智能的未来，必须要提到使用什么样的机制来实现。从计算机科学的角度来讲就是：我们不仅要谈规范，还要谈如何实现。

如果要给现有的人工智能开发方向分类，可以根据其对生物特质的忠诚程度，看人工智能的运作是否模仿了生物的大脑。一个极端的开发方向是完全没有参照生物的智力，从零开始进行设计。另一个极端的开发方向则是完全模仿生物大脑的神经网络，而且模仿得细致入微。在人工智能开发的历史上，这两个极端中间出现过各种各样的学派，对生物的模仿程度各不相同。各个学派经历了兴盛和衰落，但是没有哪种学说能取得压倒性的胜利，每一种都有自成一家的理论。

在人工智能开发这一领域里，人们常常用发明飞机的过程来类比。早期试图发明飞机的人们模仿鸟类设计了能活动的双翼，但是这种方法失败了。事实证明，固定的翅膀和螺旋桨更适合人类制造的大型、沉重的飞行器。按照这个道理，人工智能的开发不应该模仿自然界，而应该遵循更适合硅晶管计算的原则。

相反的观点认为这种类比很牵强，而且生物大脑是我们唯一能找到的普通智力的范本。我们知道可以通过神经基质实现普通智能——只要我们能人工复制神经基质，就肯定能成功。这种方法虽然简单，却有很大机会能够成功，即使我们对科学技术的发展前景比较保守。

我们之后会详细讨论不依靠仿生、从零开始创造人工智能的方法，这一部分有很多值得介绍的内容。但是，我们先来分析完全依靠仿生手段创造人工智能的方法，即“全脑仿真”[注]。全脑仿真不仅是未来创造人工智能的可行手段，还可能帮助实现“意识上传”这一某些超人主义流派的重要目标。此外，全脑仿真的一些概念对哲学思想实验很有帮助。全脑仿真涉及一系列哲学问题，比如人工智能、机器意识、个人身份等等，都和本书的主题密切相关。

全脑仿真不仅是未来创造人工智能的可行手段，还可能帮助实现“意识上传”这一某些超人主义流派的重要目标。

THE TECHNOLOGICAL SINGULARITY

1. “全脑仿真”一词由神经科学家兰德尔·科恩提出。

THE TECHNOLOGICAL SINGULARITY

第二章

全脑仿真，让仿生人工智能成为可能

为大脑制造副本

全脑仿真到底是什么？简单来说，就是用非生物基质仿照某个大脑制造一个（或多个）能运作的副本。要理解具体的细节，我们先要学习一些简单的神经科学知识。脊椎动物的大脑和它身上的其他器官一样，由无数的细胞组成。大脑细胞中包含大量的神经元。神经元是一种精妙的电结构，每个神经元都能传递复杂的信号。神经元包括一个胞体、一个轴突和多个树突。简单地说，在神经元中，树突接收信号、轴突输出信号、胞体进行信号处理。

大量神经元广泛地相互连接，形成一个复杂的网络。轴突和树突呈树状结构，枝杈末端与其他神经元的轴突和树突相连。当一个神经元的轴突距离另一个神经元的树突很近的时候，就形成了一个突触。通过复杂的化学物质交换，突触将信号从一个神经元发送到另一个神经元，这样它们就可以互相交流。大脑包含的神经元数量是一个天文数字，高达800亿个。不仅中枢神经系统（大脑和脊髓）包含神经元，动物的外周神经系统也包含神经元，这些神经元将感觉信号从身体（皮肤、眼睛、胃等）带到大脑，再把运动信号从大脑带到身体的各处（例如肌肉和腺体）。

大脑的活动是由电力和化学活动构成的。神经元的活动则受到多巴胺和血清素等化学神经递质的调控。这些化学物质由专门的神经元产生。广泛、长距离的轴突投射将这些化学物质分散到大脑中。血液也可以输送调控神经的化学物质，这就是大多数神经类药物的工作原理。

大脑不仅包含神经元，还包含血管系统，它将血液输送到大脑各部分，为大脑提供能量，大脑用这些能量产生电信号。大脑还含有大量的胶质细胞。以前人们认为胶质细胞只是一种“胶水”，把所有的神经元、轴突和树突联系在一起。后来人们发现，胶质细胞似乎也有传递信号的功能，虽然比神经元的频率要低。

个体神经元的信号传递过程已经大致为人们所了解，具体细节十分复杂。简言之，神经元收集树突获得的信息，在积累到一定程度的时候，就沿着轴突释放一种脉冲。20世纪50年代时，人们就已了解这一过程。艾伦·霍奇金（Alan Hodgkin）和安德鲁·赫胥黎（Andrew Huxsley）为这一过程建立了数学模型，并因此获得了诺贝尔医学奖。

大脑具有可塑性。在生长的过程中，胎儿和婴儿脑内的神经元会发生重要重构。轴突和树突像植物的根茎一样生长，跨越遥远的距离（以神经的标准来看）建立新的连接，淘汰旧的连接。此外，在动物的一生中，神经连接时强时弱，以帮助学习和记忆。好的数学模型也应该体现这种可塑性。

以上只是简要地概述人类掌握的关于大脑的知识。我们对大脑的了解只是冰山一角。随着人类对大脑的了解越来越多，我们掌握的知识支持着这样一个假说：人类的行为是由大脑中的物理过程决定的，是由接收到的感觉信号和发出的运动信号决定的。这一假说不仅具有实践意义，也具有哲学意义。

当然，要理解人类的行为，就要观察人类作为一种动物如何与物理和社会环境互动。否则，大脑的活动就毫无意义，但是这与我们现在讨论的假说无关。简而言之，这个假说是指从我们的所看、所听、所触到我们做的事和说的话，整个过程虽然十分复杂，但是已经没有人类不懂的环节。按照这个理论，全脑仿真是可行的。

全脑仿真的三个阶段

全脑仿真可以分为三个阶段——测绘、模拟和实体化。第一个阶段是以高空间分辨率（亚微米级）测绘全脑，至少要包括整个前脑，以保证具备较高认知功能的部分，具体来说就是大脑皮质（灰质）、连接部分（白质），以及控制情绪和行动的部分（杏仁核及基质神经节）接受测绘。测绘应当准确记录每个神经元和每个突触的位置、特点、神经元级别的连接（即每个轴突和每个树突的连接）。全脑仿真就是把某个大脑在某一特定时刻的细节精确记录下来。

第二个阶段是用这份测绘图模拟出每个神经元和每个连接的实时活动。我们可以通过标准的计算神经学技术来完成模拟过程，例如使用已有的神经元行为数学公式（如前文提到的霍奇金——赫胥黎模型）。模拟大脑使用的方法和模拟天气或模拟翅膀周围空气流动的方法是差不多的。显而易见的是，即使是模拟一个小动物的大脑，也需要极强的计算能力。

全脑仿真可以分为三个阶段——测绘、模拟和实体化。

第三个阶段是将模拟的结果外化。之前我们得到的都是计算机模拟出来的复杂模型。要把电脑里运行的数字变成外在的行动，需要制造一个身体（这个身体可能是在虚拟世界中模拟出来的身体，后文会

详细叙述）。因为模拟大脑接收和发出的信号都与生物本体相似，所以这个身体在形态和机制上与生物本体越相似越好。

如果测绘和模拟都取得了成功，那么当环境给模拟大脑和生物大脑输入一样的信号时，模拟神经元和生物大脑的真实神经元的反应应该是难以有效区分的，不管是多个神经元还是单个神经元都是如此。我在这里加了一个“有效”，是因为要做到完全一致太难了。大脑是一个混乱的系统，因此从数学上说，初期极小的不同可能会给系统行为带来巨大的差异。测绘过程的微小误差和计算过程的舍入误差都可能影响模拟大脑的行为，使结果与生物本体不同。

但是，这点缺陷不会妨碍模拟的成功。如果误差足够小，最后的模拟结果与生物大脑的结果肯定是难以区分的。从观察者的角度来看，在相同的环境下，模拟大脑和生物大脑做出的决定和行为是一样的。如果生物本体是人，那么他的朋友和爱人都会承认模拟的副本和本体的行为完全一致，有相同的习惯和说话方式，还有同样的记忆。

对小鼠进行全脑仿真

对人类全脑进行仿真需要的技术非常复杂，而且从哲学上来说具有挑战性。我们之后再具体讨论。现在，我们来看看对小动物的大脑进行仿真——小鼠的全脑仿真。这就降低了哲学和技术上的难度。如何对小鼠进行全脑仿真？需要怎样的技术？我们按照全脑仿真的三个阶段进行研究。

21世纪初开始出现对小鼠进行大脑结构扫描的技术。首先，需要杀死这只幸运的（也可以说是不幸的）老鼠，将其大脑抽出。其次，将前脑切成非常薄的薄片。再次，使用电子显微镜将每个薄片测绘并且数据化。最后，通过电脑，用数据化的绘图将每个神经元、每个轴突和树突、每个突触的位置和种类进行重建——用巨大的数据库记录下原脑的细节，这样就构成了我们需要的蓝图。

这样就足以制造仿真大脑了吗？这种方式记录的只是大脑在某个时间点的情况——组成部分的形态、如何排列、如何相互连接。但是这幅图不足以告诉我们这些部分是如何运动和互动的。扫描的空间分辨率越高，越能准确地记录神经元的微结构。对神经元微结构记录得越详细，越容易在电脑上使用数学模型重建神经元的动作。但是，即使是高分辨率的扫描也不能取得模型需要的所有参数，例如某个突触连接的强度。而如果没有取得全部参数，即使有数学模型也无法在计算机上模拟大脑。

但是，如果能取得神经元电活动的记录，即使结构扫描的分辨率不够高也能够弥补。其中一个方法（也是21世纪初的一项技术）就是使用转基因老鼠。这种老鼠的神经元会产生一种染料，在神经元发出

信号的时候，染料会发出荧光。向大脑皮层照射强光，用普通光学显微镜就能记录下神经元的每个活动（当然这要在杀死小鼠，将它的大脑切片前进行）。之后，可以使用自动化技术把缺少的参数加入模型，记录下来的数据就完整了。

这种扫描和记录的技术是非常有前景的。但是，小鼠的大脑包含7000万个神经元，每个神经元都有几千个突触连接，人脑则有超过800亿个神经元以及几十万亿个突触。从计算的角度来看，给大脑切片扫描的工作量过于庞大，即使有摩尔定律也很难实现。之前描述的荧光显微法也有缺陷。虽然空间像素很高，可以监测单个神经元，但是时间分辨率低，不能监测单个脉冲。所幸由于生物技术和纳米技术的迅速发展，其他测绘大脑的方法正在出现。我们可以先研究其中的一两个。

刚才我们谈到的方法是基因工程的一种应用，现在再来看另一种。假设我们可以改变小鼠的基因，使小鼠脑内的每个神经元都带有一个“DNA条形码”。在每个神经元都有了条形码以后，给小鼠的大脑感染一种无害的病毒。病毒被设计为在跨越突触时会携带一些遗传物质，将突触一端的神经元的条形码与另一端的神经元的条形码重组。这样会产生一组新的条形码，显示出两个神经元之间的联系。

这样，小鼠的大脑就成为一个“库”，储存着几十亿个神经元之间的连接关系。我们需要做的就是凭借DNA测序技术抽取这些数据。从数据和计算的角度来讲，这种方法比亚微米成像和成像处理少了复杂昂贵的中间步骤。这种方法的瓶颈是DNA测序的速度和成本，但是随着近年来人类基因工程的进步，这方面一直在经历飞跃。

这是一种很有前景的方法，但是和之前介绍的切片扫描法一样，不能提供大脑仿真需要的全部信息。这种方法可以表现结构，不能表现功能。这时就需要使用纳米技术了。纳米技术可以帮助测绘小鼠的神经元活动，弥补蓝图的不足。不管是生物技术还是纳米技术，都是

在探索数量巨大的超小物体。生物技术探索的是生物领域的超小物体——病毒、病菌、DNA链等。这种方法也适用于非生物超小物体。纳米技术关注的就是这类非生物超小物体的构成。这些物体的大小一般是几十纳米，也就是一米的几百亿分之一。

纳米技术的应用十分广泛，很多都与本书的主题相关，但是现在我们先着重研究如何使用纳米技术进行全脑测绘。神经元的胞体虽然只有一米的几百万分之一那么大，但是从纳米的微观角度来看，已经很大了。我们可以想象制造纳米级机器人，它们可以在大脑的血管中游来游去，每个都像贝壳一样粘在神经元的薄膜或者突触附近。它们会一直附着在上面，感受神经元蕴含的涌动，并将每个脉冲的信息实时传送出去。我们在大脑皮质表面放置微观级的接收站。这些接收站的工作就是从无数个“神经贝壳”那里接收信息，然后把信息向外界广播，由神经科学家收集起来。

虽然这些还只是设想，但在未来是有可能实现的。本书的目的不是预测科技的未来或科技设想实现的时间，而是介绍未来可行的方法以及可能出现的结果。为小鼠大脑提供一份足够详尽的蓝图以帮助进行仿真，这一设想的阻碍不是理论上的，而是技术上的。而且，这些阻碍是可以突破的，将生物技术和纳米技术结合就有可能实现。实现这一设想可能需要10年，也可能需要50年。但是从历史的角度来说，即使一个世纪也不过是短短一瞬。

还有另外一种可能，这种可能对扫描技术的要求比较低，但是需要科学进一步发展。我们在考虑是否可以复制某种成年动物的大脑，如果复制成功的话，复制品和生物本体应该是无法区别的。复制品应当与本体的行为方式、爱好完全一致，但这需要非常细致的扫描。假设我们用最先进的技术扫描大量刚出生的小鼠。然后，将所有的数据整合，再用尽可能多的小鼠大脑数据进行控制，最后，我们可以得出新生小鼠大脑的平均数据模型。

有了这样一个数据模型，就很容易生成关于某只特定小鼠大脑的具体描述，从神经元到神经元，从突触到突触。每只小鼠的大脑都有微小的不同，但是都符合总体的统计数据。这些描述不是来自真正存在的小鼠，而是由电脑生成的。但是，只要准备的数据足够多，电脑生成的小鼠大脑足以用来进行模拟和实体化。

游戏厂商助力神经模拟技术

在获得大脑的详细描述信息后，我们就可以开始准备模拟了。模拟的基质有几种选择，其中既包括传统数字计算机模拟（通过专门的软件），也包括化学或生物计算机模拟。最传统的模拟方法中使用的计算机和我们日常办公使用的计算机是一样的。只要有管理变量的公式，传统的计算机也可以用来模拟各种变量的变化——当然可能只能模拟一小部分。我们可以使用各种公式来模拟神经元的电属性和化学属性，例如前文介绍的霍奇金——赫胥黎模型。

现在的任务不是模拟一个神经元，而是模拟多个相互连接的神经元。所以，我们要面对的是很多变量，每个都依赖相关的公式。我们的任务就是同时模拟这些变量。但是，普通的电脑每次只能处理一个任务，怎么办呢？好在神经元的变化速度很慢。即使是在兴奋的状态下，一般的神经元每毫秒也只发出一次脉冲。主频3GHz（吉赫）的普通电脑可以在神经元发出两次脉冲的间隔进行超过1 000万次计算。所以，电脑可以同时计算很多神经元的活动。在模拟过程中，电脑用一毫秒的很短一部分来模拟1号神经元，再用一毫秒的很短一部分来模拟2号神经元，这样就可以同时模拟几十万个神经元。

但是，即使是小鼠的大脑也包含几千万个神经元，要同时准确地模拟这些神经元需要强大的计算能力。虽然处理器的速度在20世纪80~90年代高速增长，但是这个势头在21世纪放慢了。即使是最快的串行处理器也无法模拟小鼠脑中所有的神经元。好在我们可以进行平行处理。我们不必让一个处理器处理所有的任务，而是让多个处理器同时运行，每个处理几万个神经元的活动。1 000个工人一起干活，可以在一星期内盖起一座大楼。但是如果让一个工人来干活，就要花一辈

子。一个高速处理器可能无法模拟一个大脑，但是多个低速处理器却可以实时模拟出来。

其实大脑也是一种平行运算系统。每个神经元可以被看作一个独立的微型信息处理器。树突是这个微型信息处理器的信息来源，它用一系列物理特征来存储记忆，例如神经膜电位和突触的强度。神经元有一种“计算”功能，可以将树突“输入”的信息和神经元本身的“记忆”结合并发送给轴突。从功能角度来讲，大脑的基质就是几百万个微小的处理器在同时工作。

如果我们研究神经元的物理和化学内涵[注]，就会发现平行运算的比喻并不恰当。但是这个比喻可以让我们理解大脑的工作原理，即大量的微小物体也可以完成很重要的工作。为了模拟大脑，我们要学习这种工作方式——当然是利用不同的基质。这也预示着，2015年左右的超级电脑将大量使用平行运算。同时，不仅平行运算能使用的处理器数量不断增加，每个处理器的价格也在不断下降。这是符合摩尔定律的一种重要变化。

这种技术上的改变要感谢电脑游戏玩家。玩家们总是追求更完美的游戏体验，这就促使厂商们争相生产价格低、性能高的图形处理器（GPU）。图形处理器本来是为了处理高像素的画面，后来演变成了多用途的平行处理机。随着这种电脑的性能增强、成本下降，其他需要大量平行运算的领域也开始使用图形处理器，例如为核反应或者天气建立模型。截至2012年，世界上最强大的电脑是克雷公司的泰坦超级电脑，这种电脑使用了18 688个图形处理器，拥有了强大的平行处理能力。

---

1. 从数学角度讲，神经元的物理性能不能在传统数学计算机上表现出来，因为神经元是模拟量（所以上一段的“计算”加了引号）。

关于全脑运算的大胆构想

不久之后，最强大的电脑将可以模拟小鼠的全脑，但是要满足两个前提条件：第一，成功模拟需要的物理细节的精细度不是很高；第二，我们拥有的小鼠全脑蓝图足够详细。我们之前研究了如何在技术上满足第二个条件。但是，第一个条件能否实现还不能确定。我们能否不考虑大脑的突触传递、脑胶质细胞、树突和轴突的形状等问题，只把神经元简单地看作数学对象？如果不需要考虑这些方面，全脑仿真需要的计算量就会大大减少。

神经科学还没能回答这一问题。即使真的可以这样做，从模拟小鼠的大脑跨越到模拟人类的大脑（以及人类水平的智能），难度还是非常大的。工程师不仅要实现足够的浮点运算，还要让设备尽量不占空间和电力。人脑平均体积为1 235立方厘米，耗能只有20瓦，而2013年世界上最强大的电脑“天河二号”的耗能高达2 400万瓦，被安放在占地720平方米的房屋里。但是，即使做最保守的估计，“天河二号”能提供的计算能力与模拟人脑所需要的计算能力相比也还是相差甚远。所以，虽然数字计算机可以进行大量的平行处理，我们还是需要考虑除了依靠传统数字计算机之外，还有什么方法可以进行全脑仿真，实现人类水平人工智能。

一种可能的方法是神经形态硬件。神经形态硬件与传统计算技术不同，意在让硬件尽量与大脑的湿件[注]相似。传统数字硬件为模仿神经元在几微秒中的膜电位变化要进行几百次浮点运算。这个过程需要晶体管开关几千次，每次都要消耗电力（并且排出热量）。膜电位表现为一个二进制数字，它的变化是不连续的，而实际物理量是在连续变化的。神经形态方法摈弃了这些数字工具，使用了模拟神经元本身

结构的原件。这些原件能够模仿膜电位，具有一定的电量，并且不断变化。在节能方面，这种方法比数字电脑好得多。

之前，我们研究了全脑测绘的一些可能的方法。比如当代技术的延伸（将大脑切片扫描）、看起来可行的新兴技术（DNA编码）或者技术上可能但是距离实现比较遥远的技术（神经机器人）。对于神经模拟技术，我们也可以研究相应的可能性。我们已经研究了传统数字技术计算机进行平行运算的方法，并简单介绍了神经形态硬件。神经形态硬件可以用来模拟数量比较少的神经元，但是需要大幅提升它的效能。

除此之外，还有什么更大胆的构想呢？人们对于量子计算有过很多推测。但是，模拟大脑内的大量神经元不是量子计算的优势。叠加等量子效应可以用来帮助解决棘手的搜索问题。(注)量子计算主要用于一些需要真正大型平行处理的项目。我们需要的是一种能够突破摩尔定律的硬件，需要打破传统硬件的一些物理限制，比如光的速度、原子的大小、改写一个字节需要的能量等等。

一种可能的硬件是量子元胞自动机。虽然使用了“量子”这个词，但是量子元胞自动机并不是量子计算机。量子点是纳米级的半导体装置，作用类似晶体管，可以快速地转换状态，而且能耗非常低。四个量子点以正方形排列可以组成一个量子点电池，量子点电池可以储存一字节的信息。量子点电池可以在网格上组成逻辑门和其他阵型。这些是数字电子的基本元素，可以组成微型处理器。

相较于传统硅技术（互补金属氧化物半导体），量子元胞自动机的优势在于在同样大小的区域里，前者能放置的开关装置比后者多得多，突破了物理上的限制，而且耗电量低很多。但是把量子元胞自动机应用到生活中，还需要几十年的时间。近期，半导体工业可能会保留传统的处理器设计。一些创新的领域可能会使用3D（三维）晶体

管，这样摩尔定律仍然适用。现在使用的2D（二维）晶体管可能会被淘汰。此外，硅也会被淘汰，企业可能使用碳纳米管来制造更小、更节能的晶体管。

在物质总量一定的情况下，能用这些物质做多少计算是有理论极限的。“计算质”（Compotronium）是一种假想的物质。把这种物质用作计算机的介质，能使每秒计算量达到最大。物理学家塞思·劳埃德计算得出：质量为1公斤，体积为1升，以计算质为介质的计算机可以在10^{31}个字节上进行每秒5.4×10^{50}次逻辑运算。这比今天的电脑多了39个数量级。目前的电子工业距离这个最大值还有很遥远的距离，这一点是毫无疑问的。

也许我们永远无法实现这样的运算能力，但是只要拥有这种运算能力的很小一部分，就足够模拟人类大脑了。人类大脑的体积不过1升多，而且能耗只有20瓦（这简直让人惊讶）。不管是依靠模拟大量神经元还是用仿生的方法实现人工智能，如果出现更强大的电脑，超级人工智能就更有可能实现。

---

1. 湿件：人类的大脑，与软件、硬件相对应。——译者注

2. 物理学家罗杰·彭罗斯（Roger Penrose）声称人脑的意识和智能依赖一系列的量子运动。如果他是对的，那么使用传统计算机进行彻底的全脑仿真是不可能的。但是，很少有神经科学家支持他的观点。无论如何，这个问题与现在的平行运算都关系密切。

为模拟大脑配备合成身体

让我们假设由于某种原因，测绘和模拟大脑的障碍已经被消除，人类可以足够精确地复制小鼠的前脑。仿真的最后一步就是把模拟大脑和合成身体（机器小鼠）连接起来。只有完成了这个步骤，我们才能调试模拟大脑，使模拟小鼠和真实小鼠在行为上取得一致。仿真的机器小鼠可以有很多类型，都和真实小鼠的身体差不多。合成身体越接近小鼠的本体，连接起来困难越少。因此，我们可以假设：这个机器小鼠不是安在轮子上的，而是柔软的、有肌肉骨骼系统和四只爪子的身体。机器小鼠还有一组仿生传感器——眼睛、耳朵、胡子（非常重要），这些传感器像真实小鼠的身体器官一样传递信号。

现在我们有了模拟的小鼠前脑，也有了合成的机器小鼠身体，如何把二者连接起来呢？我们不能简单地把它们插在一起，因为真实小鼠的大脑和身体是浑然一体的，并没有整齐的连接点。小鼠的神经系统从头到脚遍布整个身体，就好像河流的干流和支流流过整个森林。小鼠前脑的神经元只是比别的地方密集得多，没有从整体中割裂开来。但是按照我们的做法，前脑和身体是分开的。在这个过程中，我们丢掉了一大部分中枢神经及外周神经。中枢神经包括小脑，而小脑在运动协调方面扮演着非常重要的角色。

我们有充分的理由相信，前脑储存的是小鼠的“精华”。我们也有充分的理由相信人类的特征都储存在前脑：习惯、爱好、专业知识、记忆、性格。因此，专注于模拟前脑是正确的。但是，我们只模拟了前脑，就好像把一幅挂毯从中间剪开。现在要想把它拼回去，必须把每一根丝线都接上，才能恢复原本的图案。或者打个更糟糕的比

方，我们把半张挂毯扔掉了，现在要重新把扔掉的部分织出来，只能依靠猜测恢复原来的图案。

小鼠的身体就是被扔掉的那半张挂毯。模拟前脑是我们剩下的另半张挂毯，需要输入和输出的信息就像挂毯上被剪断的丝线，荡来荡去。不幸的是，模拟前脑输入和输出的信息没有标记应该和机器身体的哪根线连接。工程师必须自己搞明白控制某块肌肉运动的信号是大脑哪部分发出的、大脑获得的某个信号是来自哪个感觉器官的。感觉神经元在皮质的精确位置可以为我们提供线索，特别是视觉和触觉神经元，因为这些神经元是按照位置组织起来的。但是模拟小鼠实体化需要的是接线图，只有位置信息是远远不够的。

连接之所以这么困难，是因为生物本体的各个部分（前脑、其他的神经系统、身体）是一起成长起来的，它们在这个有机的过程中相互适应特性。要解决这一问题，必须把测绘的范围扩大。与其只测绘前脑，为什么不绘制一个包含中枢神经和外周神经的神经系统图呢？在模拟某只小鼠的大脑之后，我们还可以在电脑上建立这只小鼠的高分辨率3D身体模型，包括外周神经系统和肌肉骨骼系统的所有细节。为了模拟小鼠大脑，我们必须假设技术上会出现一些重要的突破。既然如此，为什么不假设我们的技术也可以模拟小鼠的整个身体呢？

另一种方法是不扫描外周神经系统和肌肉骨骼系统，在小鼠还活着的时候，通过机器学习弄清大脑的感觉运动活动与身体得到的指令之间的关系。弄清这些关系以后，就可以建造一个连接装置，把大脑生成的运动信号转换成机器小鼠的合成身体能理解的信号，并为大脑提供本体感觉[注]和触觉反馈信号。这种方法有一个好处：合成身体不一定要和生物本体一样。不过，如果仿真机器人做好后马上就要投入使用，只有非常简单的调试和校准过程，那么生物本体的特征就必须保留——对于仿真机器小鼠来说，就是要有四条腿、爪子和一个能抽

动的鼻子。因为有了巧妙的连接装置，我们不需要复制小鼠的每块肌肉及肌肉的特点。

我们还有一个强大的机器学习工具，如果善加利用，就不用完全仿制生物的身体，这就是我们已经制造好的模拟大脑。生物的大脑最擅长学习和适应。人类能学会开汽车、开飞机、开起重机和挖掘机等。对于熟练的司机、飞行员和操作员来说，机器已经成了他们身体的一部分。此外，那些不幸因意外致残的人也展现出了惊人的适应能力，他们能学会熟练使用轮椅、假肢等工具。模拟大脑也具有这样的能力，能很好地适应新的身体。只要仿真生物不是做好后马上投入使用，就没有必要把仿真身体做得和生物本体一模一样，感觉信号可以和原来的不同。但是，仿真生物需要一段训练或者“复健”的过程，来弥合仿真身体的差异。

我们可以把两种方法结合起来——制造一个连接行为数据的连接装置，并给仿真生物一段适应时间。这样，仿真生物可以使用的身体形态就多种多样了。为什么一定要给仿真小鼠安装小鼠的身体呢？我们可以给它安上六条腿甚至轮子。如果制造者有神经指令的数学模型，比如“向视野中央的物体移动”这个指令，那么只要仿真小鼠的合成大脑下达指令，就可以保证仿真小鼠的合成身体向它视野中央的物体移动。

> 生物的大脑最擅长学习和适应。
>
> ---
>
> THE TECHNOLOGICAL
>
> SINGULARITY

一方面，模拟大脑可以适应新的身体，另一方面，我们也可以设计让新的身体适应模拟大脑。这主要依赖假肢和脑机接口领域的进

步。当代人类使用的假肢已经不再是被动的设备，相反，它们可以自己完成复杂的动作（就好像章鱼的触角）。但是，要成功地完成动作，假肢必须学会辨认用户的意图。脑机接口领域在机器学习方面已经取得了长足的进步，可以在这个方面应用，也可以帮助进行全脑仿真。如果模拟大脑和合成身体可以相互适应，那么仿真生物适应新身体的“复健”过程将会大大缩短。

---

1. 本体感觉：身体运动器官在运动或静止时产生的感觉。——译者注

从模拟人到模拟社会

生物大脑是全身感觉环路的一个组成部分。感觉环路让大脑的指令得以传达，从而使身体在3D空间中活动。因此，为了发挥作用，模拟的动物大脑也应该是感觉环路的一部分。模拟大脑必须像真实大脑一样有信号的输入和输出，所以要给虚拟大脑加上身体。一种方法是将模拟大脑与一个真实存在的机器身体连接。另外一种方法就是制造一个模拟身体，同时建立一个模拟的环境，将身体置于其中。模拟小鼠的大脑可以连接模拟小鼠身体（有模拟的爪子、胡须、皮毛等）。模拟小鼠周围的模拟环境则有模拟草地、模拟篱笆、模拟奶酪等。这些模拟环境的分辨率很高，模拟小鼠的感觉运动系统难以区分模拟环境和真实环境。

这种方法需要的技术已经非常成熟了。我们要再次感谢电子游戏产业的巨大利润。因为玩家们对游戏逼真程度的要求越来越高，所以游戏开发者研发的游戏物理引擎越来越高级，能够模拟真实物体在虚拟世界里的行为。物理引擎能表现出虚拟世界里物体的位置和方向以及它们如何运动、相互碰撞，也考虑到了重力、摩擦等因素。游戏开发商做这些工作，是为了让玩家能从角色的视角（也可能是从角色身后的视角）看到游戏中的物体。而在虚拟实体化中，物理引擎的作用是向模拟大脑提供和接收真实的信息。

不管是为了游戏开发还是虚拟实体化，工程设计方面的挑战都是一样的。坚硬的物体比较容易模拟，柔软或者有弹性的物体，例如肌肉、草叶，就比较困难了。有些特殊物质，比如烟雾和尘土，则更加复杂。但是，图形专家早已攻克了这些难题。如果模拟生物是一种社会动物，那么虚拟环境就必须具备一些更复杂的元素。这些元素的设

计可能很粗糙，比如现在电子游戏里所谓的人工智能，行为模式非常简单。我们也可以设定模拟人物为现实世界中人的化身。在实现了强人工智能以后，还可以设定虚拟世界中的人物为其他人工智能。

最后一种设定让我们认识到未来可以创造整个虚拟社会，让人工智能“住在”虚拟环境中。这个虚拟社会中的人工智能不受生物需求的限制，不需要争夺食物和水等资源。这样一来，一些真实社会中人类受湿件所限做不到的事，虚拟社会中的人工智能却可以做到。例如，当计算能力足够强大的时候，虚拟世界可以超高速运转。真实世界中的一微秒在虚拟世界中可能只是0.1微秒。

如果“住在”虚拟世界的人工智能努力提升自我，或者创造出更聪明的后代，它们进步的速度将会比真实世界中的人类快得多。如果能把它们在虚拟世界中创造的技术进步反馈给真实世界，帮助提升虚拟世界所依赖的计算基质，那么虚拟世界中人工智能进步的速度又会提高。最终，这个过程会带来类似奇点的爆炸性技术进步，而其结果是难以预测的。

把人工智能的“精灵”从瓶子里放出来

之前我们畅想的都是未来的可能，现在我们再看看短期内可能出现的情况。全脑仿真是实现强人工智能的各种可能性中，最忠实于生物本体的方法。但是，全脑仿真具有非常重大的意义。它意味着在近期，人类有希望创造至少一种人工智能（虽然只是小鼠水平）。即使是在最保守的哲学、科学和技术假设下，这个理想也是可以实现的。

我们所说的这些保守的假设包括：第一，任何动物的智慧行为都是受大脑支配的，这个过程受物理规律的制约；第二，要让仿真生物具有无差别行为，对仿真动物的物理细节的要求不是特别高；第三，已有的测绘和计算技术将在较短时间内得到大幅提高（制造仿真小鼠需要将现有技术提高2~3个数量级）。对于大部分人来说，“今生之内”及“子女今生之内”的未来才是他们关注的未来。

第一个假设是大多数人都能接受的哲学命题。第二个假设就涉及科学问题了。如果第二个假设成立，我们就不必模拟个体生物大脑的胶质细胞，生物大脑的连续性（而不是离散性）不会影响我们的模拟过程。此外，我们也可以完全忽略量子效应。关于第三个假设，如果我们考虑的是小鼠大脑仿真，那么第三个假设在计算能力方面是真的，在大脑测绘技术方面是合理的。因此，我们不难得出这个结论：小鼠水平的人工智能不仅是可能的，而且会在近期出现。

一旦我们实现了小鼠全脑的仿真，就很有理由相信，人类水平人工智能已经离我们不远了。过渡到人工智能有几种方法，最简单的就是升级仿真过程，将之应用于人脑。从工程上来说，难度可能很大，但是理论上已经成熟了。不过，相关技术真能发展那么快吗？例如，

摩尔定律不会永远适用于计算机处理能力和储存能力。小鼠大脑仿真和人类大脑仿真之间相差三个数量级，技术发展到中间水平时就可能停止了。

几十亿个低能耗、纳米级别的原件组装在一起，能够产生具有人类水平智能的设备。这一点是我们确信的，因为我们的大脑就是最好的例子。大自然做得到，人类也应该能够做到。如果只追求神经元的数量，我们最终会拥有大自然创造大脑的能力。如果没有别的办法，就把合成生物学和纳米技术结合起来。但是，要满足人类全脑仿真的计算要求，可能还需要一系列技术上的突破。从这个角度来讲，只是把小鼠大脑仿真升级为人脑仿真，可能不足以实现人类水平人工智能。

但是，要实现人类水平人工智能，不是只有全脑仿真这一条路可以走。我们可以将小鼠大脑的一些部分进行认知上的提高。例如，在大脑中最重要的部分，如额叶前部皮质和海马体，增加神经元的数量。还有一种更可行的方法，就是通过小鼠大脑仿真这个重要的研究工具，更好地理解脊椎动物的认知能力是如何实现的，然后再将新发现的理论知识应用到大脑仿真上，保留小鼠大脑仿真的核心部分，将一些重要的神经升级（例如加入认知型假肢[注]）。

从这个角度来讲，小鼠大脑仿真仍然是实现人类水平人工智能的加速器。和粒子加速器在物理学中的作用一样，小鼠大脑仿真使原来不可能的实验得以进行。例如，我们可以在严格控制的条件下观察合成小鼠大脑的活动和行为，然后对实验做一点调整（例如对合成小鼠的大脑做一点调整），把同样实验重新做一次，观察其变化。这种实验可以对小鼠的大脑进行“逆向操作”，让我们更好地理解其大脑运作的基本原理，为其设计和制造认知型假肢。

但是，这样就足以让我们向人类水平人工智能前进了吗？我们是否需要其他的东西？例如，给仿生小鼠加上语言功能就不仅仅是在某些部分增加神经元的数量的问题了。也许小型脊椎动物的大脑里没有这个功能，也许进化使人类大脑发生了质的变化，产生了语言的基石——符号表示、组合句法、语义成分。

如果真是这样，那么我们把小鼠大脑研究得再透彻也是不够的。从小鼠大脑仿真到人类水平人工智能的飞跃没那么简单。但是，我们不应该忘记，在神经工程师努力的同时，神经科学家也在努力破解大脑之谜。即使我们没有办法完全模拟人类大脑，也会有越来越强大的工具来测绘大脑的结构和活动。理解语言的神经基础是神经科学的一个主要目的。所以，当神经工程师可以制造出仿真小鼠大脑时，神经科学家已经可以给小鼠水平的仿真大脑加上认知型假肢，使之具有语言能力。

从很多方面来讲，仿真小鼠大脑都是通向人类水平人工智能的敲门砖。有些人认为，一旦实现了人类水平人工智能，我们就必然会进一步创造出超人类水平人工智能。在合成基质上创造出来的人工智能比生物大脑更容易加工提高，因为生物大脑受到各种各样的限制（速度低、离不开新陈代谢和睡眠等）。此外，新研发出来的人类水平人工智能也可以用来研发超级智能（甚至是让多个人工智能一起工作），在这样的循环中，人工智能水平提高的速度会大大加快，产生最终的智能爆炸，带来不可预估的影响。换句话说，一旦我们通过全脑仿真实现了小鼠水平的人工智能，精灵就从瓶子里被解放出来了。

---

1. 认知型假肢：帮助提高记忆力等思维能力的植入设备。

THE TECHNOLOGICAL SINGULARITY

第三章

无缝对接，人工智能走出实验室

Siri，揭开人工智能的神秘面纱

我们已经讨论了很多受人脑启发以实现人类水平人工智能的途径，特别是大脑仿真。但是，人工智能领域具有丰富的多样性，生物形态智能可能只占其中一小部分。那么，这一领域还有哪些可能性？这是一个很重要的问题，因为人工智能的形成方式会决定其行为，也会决定我们预测或控制它的能力。

如果认为其他类型的人工智能具有与人类类似的目的和动机，那就大错特错了。此外，由于构造不同，单一或集体人工智能实现目标（只要这个概念还有意义）的方式如同“外星智能”和卡斯帕罗夫下棋时的思维过程般神秘。如果一个人工智能是另一个人工智能的产物，或者它进行了自我修正或仿真进化，其隐含的神秘性就会更强。

到底是什么样的设计或建造方法会或多或少地导致人工智能无法预测、难以控制？我们对人工智能领域的可能性了解得越充分，越能做好准备，解决问题，降低制造出“错误”人工智能以及失控的风险。让我们来看看目前人工智能技术的一些例子。我们能否从这些系统中看出强人工智能的端倪？能否通过改善、拓展这些系统来实现强人工智能？其中是否缺失了什么重要的东西，需要在人工智能技术真正起飞之前予以补充？

我们先来研究一个无实体人工智能的应用：个人助理。在第二章里，我们强调了实体化的重要性，但是与我们的文化类似的诸多虚构人工智能却是没有实体的。想想《2001：太空漫游》（2001:A Space Odyssey）中误入歧途的机器人HAL。从某种程度上讲，小说中的宇宙飞船可以被看作HAL的实体。它有非常清楚的空间位置，有传感器和传

动器，可以持续与环境互动。但是在同名电影中，我们得以一窥HAL在地球实验室中的早年“生活”，观众也多少相信了它的智力与宇宙飞船无关。我们愿意搁置自己的疑虑就说明，无实体人工智能在概念上是可行的。但实际上是否可行？我们离实现它又有多远？

苹果的Siri和谷歌的Google Now展现出语音识别在过去数十年中的长足发展。它们不用事先就用户的声音接受训练，即使存在背景噪声且用户口音各不相同，它们也能够将普通语音转换为文本。有意思的是，语音识别常常不是在用户终端上完成的。原始声音文件通过互联网传输到企业处理中心，完成语音识别，生成相应的文本文档。这些无实体应用不通过感觉运动与环境相互作用，其处理和存储也都分散在云端。这是不是使这些应用显得更“无实体化”了？其实并非如此。我们可以设想一个完全实体化的机器人系统，处理则在系统外的云端进行。这还是值得我们关注的。

和将原始声音数据转化为文本一样，个人助理试图“理解”用户让它去寻找什么或对它做了何种指示。这本身就是一大挑战，即使语音文件能够完美地转化为文本。但是，如果有了巨大的样本数据库，并在此基础上建立各种语音的统计模型，事情就简单多了。只要有了问题或者命令的开头，系统就会预测出后文会是怎么样的。这种预测还可以进一步反馈到语音识别阶段，不断改善其性能，使其在有噪声或声音模糊的时候补出难以识别的信息。

用户说完话之后，系统就可以决定如何回答。用户是搜索信息还是发出指令？假设是搜索信息。是搜索与用户有关的特定信息（比如会议时间，朋友的电话号码），还是要了解普通知识？如果要了解普通知识，系统可以调动整个互联网上的资源寻找答案，然后将答案文本用已经使用多年的语音合成技术（语音合成要比语音识别容易得多）转化为语音文档，通过用户终端扬声器反馈语音回答。

这令人印象非常深刻。诚然，目前数字个人助理的对话能力还是相当有限的，你不会请它去参加聚餐。但是一代人以前，这种技术足以让天真的用户相信科幻小说里描绘的人工智能。今天，即使我们了解个人助理的工作原理，仍然会觉得它非常神秘。我们到底忽略了什么？怎样将这种对智能仿真的神秘感转变为对其坚定的信念？

让机器了解真实世界

数字个人助理的一个不可忽视的缺点就是它们并不了解真实世界。例如，它们对物体和空间关系缺乏常识性的理解，会被一些简单却意外的问题难住。这些问题的直接答案在互联网上是找不到的。比如“如果你抓住老鼠尾巴把老鼠拎起来，老鼠的鼻子和耳朵哪个离地面更近？”即使小孩也可以轻松回答这个问题。她可能从来没有抓住老鼠尾巴把老鼠拎起来，也没有看过倒吊老鼠的照片。但是人类有情境视觉化和预测行为后果的一般能力，能够处理他们未曾遇到过的问题。

日常物理是人类（和其他动物）掌握得非常好的领域，我们对这一领域基本原则的理解让我们能够解决新问题。日常心理学也是如此：人类不会像无生命物体那样行动，他们有信仰、欲望和动机。人类理解日常心理学的规则，并且利用自己的理解制订计划，进行沟通，甚至欺骗。人类对这两个领域（日常物理和日常心理学）的理解深度由人类对物体或他人思维的基本抽象概念决定。

尽管人类这些能力背后的神经机制还没有被充分认识，但我们完全可以假设，这些能力一部分是天生的，一部分是进化的结果。在原始人的生活中，物体和他人总是很突出，于是原始人选择了通用机制来应对这些事物。所以即使婴儿刚出生时并没有物体的概念，如果大脑后来仍不能把握这一概念，也会显得非常奇怪。人类了不起的地方在于能够掌握进化史上从没有出现过的，像物体或他人思维一样抽象的全新概念，比如整数或金钱。

机器怎样才能获得相同的一般能力——掌握重要的常识，如日常物理和心理学以及全新抽象概念？当然，一种方法就是复制生物大脑，我们已经相当详细地讨论了这个问题。此外，还存在其他几种可能。从日常物理角度来讲，一种可能就是采用电子游戏中使用的物理引擎，我们已经在虚拟实体化中简单讨论过这一点了。物理引擎可以对任何给定物体的组合（比如老鼠的各个部位）及其动态进行仿真。

另一种方法就是建立一个用形式语言描述日常物理法则，对万物进行逻辑推理的系统。比如，这个系统里可能有这样一句话来说明规律：没有支撑的物体常常会掉落，脆弱物体落地后常常会破碎。这就可以推理出，如果红酒杯从桌上滚落，就会破碎。以逻辑为基础的方法同样适用于其他领域，如日常心理学。与物理引擎相比，它拥有额外的优势：可以接受不完整信息，比如不知道桌子和红酒杯的确切形状。

但是，物理引擎和逻辑推理都依赖人类设计师提供的概念框架。机器人周围的环境（物体表面）可以通过来回移动、收集传感器信息（来自摄像头、触觉传感器等），把数据转化成适合处理的形式。但是物体这一对整个系统预测能力非常重要的概念，并不是通过与环境互动产生的，而是从一开始就预置于系统内。这对于某些具有普遍重要性的知识领域（如日常物理）是可以接受的，但是真正的一般智能要能够自己发现（或发明）抽象概念，以应对未知世界。

建立模型，学会处理不确定性

我们来研究机器学习这一话题。人工智能出现伊始，机器学习就是非常活跃的子领域。但是直到进入21世纪后，随着计算和存储的能力不断增强，理论发展和新学习算法出现，这一学科才取得了长足进步。这带来了新的商业应用，如线上营销，机器学习可以更好地描绘消费者特征，更有针对性地向消费者推荐产品。有了消费者购买习惯和浏览习惯的巨大数据库，机器学习系统可以此为基础建立消费者行为的统计模型。有了这一模型，系统就可以根据消费者的历次购买行为和网页浏览情况预测其偏好。

通常来说，机器学习涉及建立模型和解释已有数据，也可用于对更多数据进行预测。例如，我列出一组数字：5,10,15,20，然后请你猜一猜接下来是什么数字。你很可能会做出这种假设：数列中每个数字都在前一个的基础上加5，因此下一个数字是25。如果数据来自真实世界，它很可能是杂乱无章的。所以，机器学习算法必须能够处理不确定性。假设可移动机器人是静态的，但是一个巨大物体正在接近它。机器人获得了一系列表明物体与自己距离的传感器读数： 24.9厘米，20.1厘米，15.1厘米，9.9厘米。它可以做出这一假设：每一次传感器发来读数，物体与自己的距离都缩短5厘米，那么下一个读数将会是5厘米，上下浮动10%左右。这时就该采取躲避行动了！

在这些例子中，发现潜在规律很容易。但是，如果每个数据不只包含一个数字，而是包含1 000个数字。从这样的高维数据中发现规律，建立模型并进行预测就困难多了——可不止困难1 000倍，这叫作维数灾难。令人欣慰的是，数据表现出的已知统计规律可以降低维数灾难的难度。

例如，如果我们讨论的数据是视频中的一系列帧，那么就会有这样的统计趋势：（1）任何给定帧的单一像素值都与邻近帧的单一像素值接近；（2）连续帧中的同样像素具有类似值。

这样的统计规律通常表现出提供这些数据的世界本身的基本结构。对于装有摄像头的移动机器人来说，世界表现出一种“光滑性”，充满了各种物体，物体表面有大量连续的色块，偶尔会有些断裂的边缘。虽然可以在设计时就将世界结构的若干假设预置于学习系统内（如其3D空间特征和世界上有众多物体），但是世界包含的物体及物体行为还需要不断探索。

所以，通过建立世界模型学习如何预测输入数据这一任务本身，意味着想办法压缩数据，降低其维数。例如，通过“动物”、“树”和“人”这样的概念与类别来重新描述世界（这也是语言交流的又一有益基础）。但是，高维数的传感数据不能被直接压缩成这样的类别，需要采用结构性方法，先提取低层次特征。在制作出低层次视觉特征循环表之后，算法就可以学习低层次视觉特征是如何形成更高层次特征的。这种多层次手段就是深入学习的标志。

如果数据来自真实世界，它很可能是杂乱无章的。所以，机器学习算法必须要能够处理不确定性。

THE TECHNOLOGICAL

SINGULARITY

假设学习算法要处理一个包含人脸图像的大型图像数据库。人脸可以表现为一系列明暗变化明显的色块的特殊组合。这些可能会与我们称之为眼睛、鼻子和嘴巴的特征粗略对应，也可能无法对应。机器不受人类语言类别的限制。机器挑选出的低层次、在统计学上显著的

视觉特征可能并不符合直接的语言描述（实际上，生物大脑的视觉认知也是如此，尽管人类大脑受到语言自上而下的影响）。

在了解数据低层次的统计特征（反复出现的小规模的视觉符号）之后，学习算法可以发现这些符号的某些特定组合常常出现。其中一种组合可以与所谓的人脸对应，另一种组合（胡须、皮毛、尖耳朵等）可能是人类所说的猫。由于猫常常被小孩抱在怀里，算法可能还会选出儿童——猫这一组合。这再一次表明，机器不受人类概念类别的限制，只遵从数据统计。

到目前为止，一切都还不错。我们看到机器学习算法是如何处理静态数据的，不过我们最感兴趣的是世界的动态变化。我们已经假设了可以在静态图像数据库中识别物体类别的系统，那么视频档案怎么办？毕竟实体化的学习系统要能够处理动态图像（也就是源源不断输入的传感数据）才具备预测能力。对于要带来满意度、实现目标的人工智能来说，只有在猫表现出某种独特行为时，才有必要将猫从周围背景中识别出来，特别是当这一行为与人工智能的动机和目标相关时（假设人工智能是一只老鼠）。

例如，如果我们的学习算法掌握了“线”和“猫”这样的类别，它很快就会发现猫常常玩线球。但是我们也不能错误地认为，机器学习算法会用人类语言来表述这一规律。机器学习算法的表达体现为数据结构内的一系列参数，这些数据结构捕获了经常出现的视觉特征运动的统计规律，视觉特征同样由数学方法表达。但是，对构造合理的机器和了解这一事实的人类来说，结果其实是一样的。比如，可以帮助人工智能制订良好计划，将猫诱骗到篮子中，再带去兽医那里。

大数据带来人工智能

我们来整理一下思路。我们已经讨论了算法可以学会世界的统计规律，可以发现多模式未分类数据中物体的类别结构和行为，可以利用这些类别将数据转化成数学表达再进行预测。我们可以看到，机器学习算法是一项很有用的技术，但是它能带我们往人工智能的方向走多远呢？

假设我们按照下文所述的方法构建人工智能。我们描述的学习算法可以在互联网上像搜索引擎一样搜索，对数十亿图像和数亿视频进行统计分析。人类在日常生活中产生了大量的多媒体数据。只要有网络，任何人都可以获取这些数据。在网上可以找到长颈鹿交配、飞机翻跟头、印度人种土豆、中国女孩修自行车、打仗、董事会议、建筑工地或无所事事的可爱猫咪的视频。无论什么，都有人拍下来放到网上。

由于社交网络的众包，公共数据的数量仍在迅速增加，而且很多视频不再只是原始传感数据。图片和视频短片通常配有地点、时间和日期信息，越来越多的视频还对物体和事件添加标签。随着越来越多的常用物品（垃圾桶、冰箱、钥匙链等）与互联网建立起联系，我们有可能收集更丰富的世界的信息，以及居住其中的人类和其他动物行为的信息。

运用强大的机器学习算法来处理庞大的数据库，系统由此做出的预测又有多准确呢？为什么这一系统需要实体化？既然已经有了记录其他诸多实体行动的多媒体数据库，为什么系统还需要与世界进行直接互动？我们回忆一下就会发现，“让计算机拥有对普通世界的常识

性了解”一直都被视为走向强人工智能的一大挑战。无实体人工智能或许可以通过间接方式获得常识，这样的系统又与人类水平人工智能有多大差距呢？

语言方面的情况怎么样呢？语言是人类行为非常重要的一个方面。如果人工智能不能达到人类的语言能力，就无法被称为人类水平人工智能。现在的数字个人助理善于预测它们的主人要说什么，已经令人意外了。但是我们也可以轻描淡写地说，它们其实并不理解识别出的词语，听到的句子或提供的回答。它们使用的符号并不是来自与世界的互动，在被问到需要将想象力和常识结合才能回答的新问题时，这一缺陷就会凸显出来：“如果你拎起老鼠尾巴，到底是它的鼻子还是耳朵离地面近？”

是不是无论机器学习多么强大，都无法帮助我们克服这样的局限性？语言其实只是行为的另一种形式。为什么通过统计进行机器学习，不像人群运动规律或植物种植规律那样容易受到强力影响？只要提供足够的数据，进行足够的计算，机器学习就能够对相关统计数据进行建模，做出准确预测。这个人离开报刊亭后可能会去哪里？那棵树左边的叶子可能会是什么形状？这个人对那个人说的话会有什么反应？我们不能忽视的是，与今天的数字个人助理相比，我们设想的学习系统依赖一个更大的数据集，其中词语来自经验，来自与世界的实体互动——无论是间接的还是寄生性的互动。

倒吊老鼠的问题该怎么解决？人工智能要能够处理假设的、违背事实的、想象的情景，这是一项基本功能。如果模型要充分模拟世界，具有完善的预测能力，其初始设定就必须是假设情景，也就是充满想象物体的情景。模型的预测功能会完成其他任务，从数百万倒吊事物的视频和图片，包含老鼠各种姿势、活动的视频，以及数十亿可以想象到的耳鼻位置关系中，概括出规律。

数学又该怎么解决这个问题？肯定没有统计学习系统能够掌握数学能力吧？（哲学系的学生会在实证主义和理性主义之间看到类似的辩论。）当然，我们没有禁止在系统中预置各种类别和概念，比如物体、3D空间或数字的概念，但是我们尚不清楚系统需不需要这些概念。对于学习算法来说，可能只处理大量小学数学课录像，就足以让它发现数字的概念了。我们很难想象学习算法要处理的原始数据量，系统的处理方式可能也会让我们大吃一惊。

2009年，谷歌的三位计算机科学家写了一篇论文，题为“数据不合理的有效性”。这个题目指出了机器学习的一种意外现象。实际上，使用含有一万亿条目的混乱数据集处理任务（如机器翻译）更有效，只有100万条目的干净数据集则根本没用。说这种现象令人意外，是因为100万似乎已经是个很大的数目了。如果学习算法不能很好地运用含有100万样本的数据集，我们的直觉就告诉我们，这个算法可能不行。但是现在我们发现，算法需要的是更大的数据集，这是只有在计算机强大到能够存储和处理这么多数据时，我们才能了解的。

这一教训告诉我们，如果人工智能系统从一开始就根据与生物大脑完全不同的原则运转，我们感到意外也就理所当然。特别是人工智能系统处理的数据量如此庞大，处理速度如此快，靠人类的直觉根本无法把握。之后，人工智能甚至可能以我们无法完全理解的方式令人意外地解决问题。简单来说，人类水平人工智能不一定要像人。如果连人类水平人工智能都如此神秘，我们又如何能够预测和控制超级人工智能？超级人工智能可能在每个智力领域都能够与人类匹敌，甚至比我们还聪明。

用最短的时间，去最多的城市

预测能力本身并不是强人工智能的全部。对世界建模，并运用这些模型进行预测，都是实现其他目的的手段。动物智能反映在它的行为中，表现出一种目的感。动物有内驱力，比如饥饿和恐惧，可以形成有益于内驱力的目标，比如获得食物或回到家中。它通过对世界采取行动来实现目标。如果动物很聪明，它也会通过预测帮助自己实现目标。当猫看到一只老鼠消失在树桩后时，它预见到了老鼠会再次出现，于是会耐心等待。我们希望实体化强人工智能的预测能力也能够服务于其目标和内驱力。它应该表现出自己的目的感。无论是打包快递、做饭，还是进行外科手术，只有当它有目标并且能够实现目标时，我们才会觉得机器人具有通用智能。

无实体的人工智能又是什么情况呢？要拥有强人工智能，系统不只要完成预测，即使它的目的只是回答问题、提供建议。尽管自身并不能对世界采取行动，它也应该能很熟练地找到实现一系列给定目标的行动方法。比如，要求它设计可赢利的投资组合，规划大型土木工程项目，或者设计更有效的药品、更大的飞机、更快的计算机。如果其智能真的是通用的，就有可能训练它完成其中任何（或所有）任务，如同训练一个拥有智能的人。

如果连人类水平人工智能都如此神秘，我们又如何能够预测和控制超级人工智能？超级人工智能可能在每个智力领域都能够与人类匹敌，甚至比我们还聪明。

THE TECHNOLOGICAL

所以，无论机器是否实体化，要完成如此具有挑战性的任务，除了预测能力还需要什么呢？人工智能需要能够计划一系列行动，善于计划就意味着善于优化。实际上在当代强人工智能研究伊始，优化就是核心问题。不仅规划可以被看作优化，机器学习和计算机视觉的某些领域以及与人工智能有关的很多其他问题，也可以被看作优化。所以我们应该更深入地研究这个概念，让我们用一个更具体的例子——推销员出差问题来进一步解释。

假设一位旅客（或者推销员）要前往好几个城市，然后回家。她必须到达每个城市一次，最后回到出发地点。她选择到达城市的先后顺序会影响整体的旅行时间，而她不想浪费时间。假设她住在旧金山，要去纽约、波士顿和圣何塞。因为旧金山和圣何塞离得很近，和纽约、波士顿离得很远，所以如果她从旧金山去纽约，然后去圣何塞、波士顿，最后回到旧金山就不是最优选择。如果她去了纽约之后直接去波士顿，出差时间就会大大缩短。这里的挑战在于要找到最优方案，即前往不同城市的最佳顺序，也就是说，按照这个顺序会使旅途时间最短。

推销员出差问题只是优化的一个例子。一般来说，需要找到非常清晰的数学结构，使成本函数最低（或使回报函数最大）。在这个案例里，数学结构就是到访城市的顺序，成本函数就是总旅行时间。如果只是去这么几个城市，问题并不复杂。但是，像很多优化问题一样，推销员出差问题很难拓展开来。从特定的数学意义上讲，如果城市数量增加，这个问题的难度就会呈指数级增长。

这也就是说，如果城市数量很多，即使是用最快的传统计算机运行最快的算法，要在合理时间内找到最优方案也是很吃力的。尽管可能找不到最优方案，有些算法也会针对大量城市的情况找到相对较好

的方案。因为推销员出差问题并不只是由求知欲造成的，它还具有很多现实意义，通常找到一个相对较好的解决方案已经很好了。

在重新讨论强人工智能之前，让我们再来思考一个相对较好的解决方案便已足够的优化问题。假设问题的主角不是推销员，而是一只叫作Tooty的猫。Tooty不是要去不同城市出差，而是要在每次睡醒之后，去不同的觅食点找吃的（比如邻居家的厨房）。当然，从一个觅食点到另外一个觅食点需要消耗能量，它要尽可能消耗最少的能量，吃到最多的食物。令它烦恼的是，它并不一定每次都能够找到食物（邻居家的猫可能捷足先登）。但是，根据自己的经验，Tooty“知道”在每个地点找到食物的概率。

现在，Tooty的任务就是制订觅食路线图，使它的预期回报最大化，它觅食的回报是总食物摄入和消耗能量的函数。和推销员出差的例子不同，觅食并不需要到每个地点去。所以比较好的方法是跳过那些偏远的、希望不大的地方。否则这个优化任务就和推销员出差问题没有太大差别了，而且计算难度也至少与前者持平。这个问题的新特点就是不确定性。但是，无论Tooty制订一个多么完善的计划，都不能保证它会找到的食物量。如果运气不好，它可能一天什么也找不到。

但是现实中有很多不确定性。无论机器学习算法多么聪明，都不能建立每次都做出正确预测的模型。相反，如果数据有限且不完整，我们能够指望的也就是得到一个概率模型，能够预测出最可能出现的结果。有了概率模型之后，在选择行动方案时，就要根据模型选择能够使回报最大化的方案。但是，我们面对的还是非常具体的优化任务。

不确定性不会使我们超越数学和计算的边界，而是会带我们走进概率理论的数学范畴。

人工智能必须回答的三个问题

真正的猫不会像我们描述的那样。真正的Tooty不会建立食物供给的概率模型，不会到处觅食还什么都吃不到，最后回到窝里重新制订优化路线。猫和其他适应性很强的动物一样，会在觅食的时候学习，在学习的时候觅食。探索世界和利用世界的资源完全是合二为一的，这也是正确而理性的策略。因为我们会看到，类似的将机器学习与优化结合起来的策略也是强人工智能的坚实基础。

对于人工智能的研究者而言，在不同情境下尝试不同行动，看看怎么做最有效且最大化预期回报，被称为强化学习。推销员出差问题和觅食的猫都是非常具体的优化的例子。如果只解决推销员出差问题，这种算法无论多快，都无法成为强人工智能。相反，以预期回报最大化为中心的强化学习概念，并不与具体问题相绑定。实际上我们可以在这一理念基础之上，明确通用人工智能（Universal Artificial Intelligence）的具体形式。

马库斯·胡特（Marcus Hutter）首先准确提出了通用人工智能的概念，和提出通用计算的艾伦·图灵一样，为计算机科学做出了重大贡献。通用计算机是指只要有了合适的程序就可以进行任何计算的计算机。图灵的成就在于他从数学上把握了计算机的这一理论。和图灵的抽象计算设备（我们现在称之为图灵机）不同，真正的计算机受限于有限的内存。但是从理论上说，每台数字计算机都可以做任何计算。它们从图灵的数学描述中获得了通用性。

与之类似，通用人工智能无论所处世界状况如何，在收集信息后，都会选择回报最大化的行动方案。可以说这就是完美的人工智

能，它能够充分利用获得的数据做出决策。和图灵的通用计算一样，这个概念也可以用数学进行准确表达（我们这里不予详述）。和图灵的定义相同，这种数学理想在现实中也不可能实现。但它可以作为人工智能的理论上限，如同图灵的定义充当着计算这一概念的理论上限。

尽管不切实际，通用人工智能的概念却不只是数学家的玩具。从一开始，类似的概念就可以在现实中实现。但与我们的讨论关系更密切的是，胡特的数学描述暗示人工智能具有简单、一般性的架构。这一架构中有两个流程交叉进行：机器学习，旨在针对世界构建具有预见性的概率模型；根据这些模型明确预期回报最大化的优化行动。

这一架构有着非常广泛的应用。实际上，任何媒介，无论是人工还是生物智能，都可以根据这一架构进行分析。我们必须回答三个（或者三组）问题。第一，这一媒介的回报函数是什么？回答这个问题会告诉我们它将如何表现。第二，它怎样学习？处理什么数据？使用什么学习技巧？预置其中的日常知识有哪些？第三，它如何实现预期回报最大化？它做到这一点的优化技巧有多出色？它善于解决什么问题，弱点和局限性有哪些？

想想非人类动物，比如一只乌鸦能够通过反复试验学会复杂的行为，并且在解决问题时具有一定的创新性。它的回报函数是什么样的？和所有动物一样，乌鸦的回报函数是获得食物和水并躲避风险。这些看起来可能是很简单的需求，但是乌鸦获得食物的障碍很可能是人为制造的。

比如，为了测试乌鸦的认知能力，研究人员会把一条虫子放在盒子里，要想打开盒子盖，必须解开谜题。乌鸦是特别聪明的动物，可以解决简单的规划问题。但是同样的形式，可能题目会更难。比如，一只不那么幸运的乌鸦可能不得不走赢一盘象棋，才能把盖子打开——这只乌鸦肯定要挨饿了。关键在于获得食物等资源这一需求可以

被看作普遍回报函数。在复杂环境中，任何问题都可以转化为获得单一资源要面对的挑战。

关于回报函数的讨论就到这里。第二个问题是乌鸦如何学习。乌鸦与物质世界进行实体互动，世界表现为无数的物体，既有生命体，也有无生命物体，其外形和活动状况各不相同，乌鸦在这一世界中，从自己感官获得的数据中学习。它学会了当自己推、戳、啄、冲着这些物体尖叫，或者干脆不予理睬时，这些物体会有什么反应。实际上，当它这么做时，其神经基础是什么是我们没法回答的科学问题。但是，动物认知研究人员已经让我们充分了解到乌鸦这样的动物可以建立怎样的认知联系，它们能做出哪些感官识别。

在明确能使预期回报最大化的行动时，乌鸦到底有多出色？答案是乌鸦比绝大多数动物都要出色。它可以采取丰富多样的行动，包括使用工具。这就构成了先天刺激反应行为的基础，刺激反应是进化的功劳，进化使得关于世界的一些假设帮助乌鸦实现回报最大化。但是，乌鸦并不只是僵硬地对刺激做出反应。它可以发明新的行动序列来解决以前没有遇到过的问题，有时还会发明一些新的行为（如制造新工具）。这一能力的神经基础现在仍然是个谜。但是和其他非人动物相比，无论乌鸦的优化方法是什么，都具有很强的通用性。

所有这些让我们了解了很多乌鸦的能力和局限，帮助我们预测它们的行为。比如我们知道，乌鸦可能会弄倒垃圾箱，从里面寻找食物碎屑。但是我们不用担心它会破解银行账户，盗取我们的资金。要更好地了解不同类型人工智能的能力和局限，我们可以提出同样的问题。不同类型回报函数带来了哪些影响？人工智能装载了什么类型的机器学习技能？它们处理什么数据？要使人工智能的预期回报最大化，可能会使用什么样的优化算法?

人类水平人工智能，学会享受生活之美

和猩猩、狗、大象以及其他很多非人类动物一样，乌鸦的聪明常常令人惊讶，但是它们远没有人类聪明。动物级别的人工智能已经很有帮助了。具有和狗相同智商的机器人可以完成一系列重要工作，但是我们真正关心的是人类水平人工智能。我们想知道如何打造出可以在各个智力活动领域都能与人类匹敌，甚至在某些领域超过人类的人工智能。或者我们至少要具体了解这样的人工智能是怎样运转的，并以此为基础想象未来拥有这种机器的世界将会是怎样的。之后，我们可以再考虑在所有智力领域超过人类的超级人工智能的可能性。

无论我们考虑的是人类水平人工智能还是超级人工智能，我们都需要像之前那样提出同样的三个问题：它的回报函数是什么？它如何学习以及学习什么？它如何优化预期回报？但是，当我们开始这一想象力练习之前，针对智人提出这三个问题大有裨益。第一，人类的回报函数是什么？我们肯定和其他动物有着相差无几的潜在回报函数——人类需要食物和水，不喜欢疼痛，注重享受。而且人类回报函数和乌鸦一样是“通用的”：在理论上，智力挑战可以转变为获得食物的智力挑战。但是，人类能够极大修正他们的回报函数，这很重要。

很多动物都会学习把物体或事件与回报联系在一起。在“巴甫洛夫的狗”这一非常知名的实验中，通过不断用铃声和食物同时刺激狗，狗学会了将铃声和进食联系在一起。最后，即使没有食物，狗听到铃声也会分泌唾液。这有益于使预期回报最大化。在需要抢夺食物的情况下，这条狗会比不知道这种相关关系的狗更早跑到碗边，得到更多食物。但是在这种情况下，潜在回报函数依然深深植根于生物本能中，并没有真正改变。

相反，人类从孩童时代开始，各种相关关系便重叠在一起，并同时受到复杂的社会暗示和期待的调节，这可能会导致回报函数和生物本能完全脱节。甚至有人认为，我们人性本质的一部分就是超越生物需求。人可以演奏音乐、写诗、设计花园，毫无疑问，这些活动常常是为了获得金钱回报或社会地位，其动机是可以用生物需求来解释的。但有时它们确实是出于对美好生活的思考，因而其本身就成为目的，并不是觅食、避险或是其他具有进化价值事物的替代品。

这就使我们面临第二个问题：人类如何学习世界，以及人类和其他动物在学习、了解世界时是否存在差异？回答是显而易见的。由于社会、文化尤其是语言的存在，人类回报函数具有开放性。有了语言，我们才能反思自身所处的状况，就如同我们在哲学、艺术和文学中所做的那样。没有这样的反思，我们就不知道如何像现在这样超越生物需求。也是由于语言，人类能够在技术发展中进行合作，一代人的技术成果能够轻松地传递给下一代。所以除了了解日常物理、自然和社会环境，人类必须要学会语言。能够理解别人的信仰、欲望和情感状态，会使学习变得更易驾驭。

第三，人类如何使预期回报最大化？社会、文化和语言也很重要。人类的智力具有集体性。人类技术不仅是很多个人，也是很多代人劳动的成果。知识、专长和基础设施层层叠加，每一代人薪火相传。因此，为了实现社会回报最大化，个人的优化能力进一步专业化。个人的回报函数是可敬还是可鄙，个人是圣人还是罪人，其实没有什么区别。一个人在社会中使用语言，可以明白如何从他人那里得到自己想要的东西。

无论对集体还是个人，创新能力是人类优化回报策略的另一关键要素（回忆一下第一章，能否赋予计算机创造力被称为走向强人工智能的主要障碍）。建筑、写作、印刷、蒸汽机、计算机等发明都对人类健康、寿命和普遍福祉做出了极大贡献，因此在很长时间里，有效

地使回报实现了最大化。人类回报函数除了青睐身体健康和长寿等因素，还受到性选择、社会地位竞争和其他特殊生物因素的影响。结果就是出现了一些不那么实用的创造形式，比如舞蹈、宗教仪式、时尚、艺术、音乐和文学。

现在，如何使人类水平人工智能工程从一开始就走上这样的道路呢？这三个关键问题（回报函数、学习和优化的问题）能够在多大程度上揭开此类人工智能的奥秘呢？当然，如果人工智能要达到人类的程度，即使其设计构造与人类大脑完全不同，也应该基本符合我们之前所描述的那些行为方式。但是，就像我们在《数据不合理的有效性》一文中注意到的那样，人类水平人工智能不必与人类相似。只要人工智能能够在大部分智力领域达到普通人的水平，在少数领域超过普通人，就可以被称为人类水平智能。

如同我们在不同人身上看到的那样，这其中又有很多差异。有些人数学好，有些人语文好。有些人善于和人打交道，有些人善于在家里研究技术。人类水平强人工智能可能有非常大的工作存储器，或是非常善于在数据中搜索规律，但是却不能写出一部有价值的小说或创造出一种新的音乐形式（当然大部分人也不能）。但是，如果人工智能能够在每个智力活动领域与人类匹敌，甚至超过人类，又会怎样？这种超级智能机器是否会出现？它带来的后果是什么？这些问题我们将在下一章解决。

THE TECHNOLOGICAL SINGULARITY

第四章

超级智能，想象另一种可能

通往人工智能之路

现在，能够帮助人类制造人类水平或更高水平人工智能的技术已经有很多，有些是仿生技术，有些则是完全从无到有的创造。这些技术可以创造出一些基本元素，基本元素进行不同的组合，进而制造出各种各样的人工智能。为了了解未来人工智能怎样运作，我们再来回答上一章提出的三个问题：人工智能采用怎样的回报机制？人工智能怎样学习？如何实现回报最优化？

我们还可以提出一系列更富有哲学意味的问题。人工智能有道德的概念吗？如果有，人工智能应该为自己的行为负责吗？人工智能会痛苦吗？如果会，它应当享受权利吗？人工智能会给社会带来怎样的变化？会给人类带来怎样的变化？如果不限制它们的行动自由，它们会给人类世界带来怎样的改变？会对经济、社会结构以及人的本质带来怎样的影响？人工智能会带来怎样的世界？人工智能的出现是会带来乌托邦、反乌托邦，还是不会带来很大的变化？

在我们详细回答这些问题之前，先要研究一个重要的命题：一旦人类制造出了人类水平人工智能，就一定能制造出超级智能。要判断这一命题的真假，我们先要研究数字基质相较于生物基质的优势。与生物大脑不同，数字基质的大脑可以多次复制。数字基质的大脑还可以加速，这一点也与生物大脑不同。所以，如果我们能通过全脑仿真制造人类水平人工智能，那么只要运算能力足够强大，就可以制造一个以超高速运转的人类水平人工智能的社区。如果人工智能不是按照仿生方法制造，而是从无到有创造的，这个道理也适用。事实上，只要是电脑程序，都可以复制或加速。

这个道理的意义非常深远。为了理解其意义，我们可以想象一下具体的场景。假设一家知名公司决定在新兴市场开发一款高性能的摩托车。公司跟两家机动车设计公司签订了奖励合同。哪个公司设计得好，其产品就将投入生产，并且获得高额的设计费。一家设计公司雇用了传统的人类设计师。另一家设计公司是一个创业公司，他们有一系列人类水平人工智能，这些人工智能生活在虚拟社区里，可以参与这种大型项目的设计。

与生物大脑不同，数字基质的大脑可以多次复制。数字基质的大脑还可以加速，这一点也与生物大脑不同。

这个项目需要很多领域的专业知识，包括材料、引擎设计、流体力学、人体工程学等等，此外还要了解什么样的外形会受欢迎。从设计概念到制造出第一辆样车需要最优秀的人类工程师工作两年。人工智能设计公司处于劣势——他们没有雇用机动车设计专家。但是，他们拥有强大的运算能力和最先进的人工智能技术。所以，从零开始培养一支设计师队伍是没有问题的。

首先，这家公司购买了一些基础人工智能，组建了一个人工智能社区。这些人工智能默认设置为拥有20岁人类的平均经验，接受过机械工程或工业设计的研究生水平教育。这一支人工智能队伍现在还算不上机动车设计团队。另一家公司的人类设计师已经有多年设计汽车、自行车和引擎的工业设计经验。要追上对手，人工智能团队必须获得至少同样水平的经验。幸运的是，人工智能可以在虚拟社区里完成这一工作，执行一些小项目，有些是独立完成，有些是团队完成。

当然，如果在现实中进行培训，人工智能团队肯定赢不了——它们还没做好准备，人类对手就已经可以生产出样车了。但是，假设人工智能的虚拟时间比现实时间快10倍，那么10年的培训和设计任务可以简单地在12个月中完成。在项目第二年开始的时候，人工智能小组已经追上了人类小组。此外，它们还有一年的现实时间，相当于10年的虚拟时间。它们可以在10年内设计出完美的超级摩托车，而人类小组只有一年时间了。一组年轻热情、有干劲的工程师在10年里能创造出怎样的奇迹，我们可以想象一下。

在项目结束的时候，两个小组都要提交自己的设计。传统设计公司提交的是一辆不错的样车，外形时髦优雅，肯定能吸引消费者。但是人工智能团队的设计又怎么样呢？当这个团队向大家展示样车时，所有人都惊呆了。没有人见过这样的摩托车。它的外形极具颠覆性，细节更是超乎人们的想象。这款车提速非常快，不仅能保持高速行驶，还非常省油。

使用人工智能的企业获胜后，向人们公开了他们设计的秘密。他们有充裕的设计时间，开发了一系列适合摩托车制造的生物材料。他们还开发了一种燃料预处理技术，使用了化学领域从未使用过的科研成果。此外，他们开发了一种制作方法，可以使摩托车的电子元件和车身融为一体，采用一体成型工艺。这些技术都申请了专利，设计公司除了赢得设计比赛的奖金外，还可以从专利中获利。

这个小故事告诉我们，一旦实现了人类水平人工智能，超级智能也会很快出现，不需要创造另一种形式的智能，也不需要概念上的突破。即使人类水平人工智能通过最传统的方式实现（即完全依靠仿生），只靠人工智能可以超高速运行这一点就足以超越人类水平人工智能。但是，这真的是超级智能吗？人类水平人工智能能高速完成的工作，如果给予足够的时间，人类智能也能完成。

也许个人超级智能与集体超级智能有所区别。我们这个小故事讲的是集体超级智能。人工智能团队里每个个体都不符合超级人工智能的标准。但是，战胜人类智能的是集体人工智能还是个体人工智能，对于是否能建造超级人工智能的讨论并不重要。失败的那一组工程师即使知道自己是被一组人工智能而不是一个人工智能打败的，心情也不会更好。同样，如果因为人类水平人工智能的出现，人类最终进入乌托邦或反乌托邦，那么改变人类的元凶到底是不是“真正的”超级智能并不重要。

最重要的是技术到底能做什么。著名科幻小说家亚瑟・C・克拉克（Arthur C. Clarke）写过一段著名的话：“真正先进的技术在人们看来就像魔术一样。”无论人类水平人工智能是怎样实现的，一旦成功，就会带给我们像魔术一样的先进技术。因为人类水平人工智能比人脑运转得更快，之前的小故事已经阐明了这一点。但是，要了解实现人类水平人工智能可能带来的颠覆性结果，我们必须依据技术的本质考察提升人类水平人工智能的其他方法。我们会研究如何从无到有创造超级智能，但是在此之前，我们先来研究仿生人类水平人工智能。

不吃不睡不领工资的超级智能员工

在设计摩托车的案例中，人工智能小组仅因为工作速度更快，就取得了明显的优势。如果人工智能是按照大脑设计的，那么它的工作速度可以设置为比真实世界的速度更快。大脑从生物基质转移到数字基质是一种解放，在这一过程中获得的最简单明了的优势就是可以超高速运转。但是，还有很多方法可以提高仿生人工智能的工作性能。

人类员工受到各种各样的生物本能的限制。例如，人类需要吃饭、睡觉。但是，即使仿生人工智能是非常忠实地按照某个大脑进行全脑仿真制造的，也可以摆脱这些本能的束缚。真正的大脑需要血液供应葡萄糖，以保证神经元能够工作。仿真大脑就没有这样的需求，至少电脑中的模拟大脑不需要（电脑需要能源，但那是另外一个问题）。睡眠比较复杂，做梦似乎有着重要的心理功能。所以，在全脑仿真中，可能不能完全取消睡眠。但是，如果按照脊椎动物的神经系统的运作方式进行设计，可能就可以取消睡眠。

简言之，仿生人类水平人工智能不需要浪费时间寻找食物、准备食物、吃食物，也不需要像人类那么多的睡眠（全脑仿真则需要足够的睡眠）。节省下来的时间可以用来工作，这也可以带来相对于人类的优势，虽然优势没有超高速运转那么明显。人类员工肯定不会同意把吃饭和睡觉的时间都用来工作，但是设计出来的大脑可以把回报函数设计得和人类不同。对于很多公司来说，不需要吃饭睡觉、只想工作的智能奴隶简直是完美的理想员工，特别是它们还不需要薪水。

消除对食物和睡眠的需求是从生物本能解放的最直接的方法。还有很多方法可以让仿生人工智能获得优势。很多人通过摄入咖啡因这

种已经被普遍证实的方法来提高认知水平。裸盖菇素（蘑菇中提取的活性成分）等迷幻剂经常被用来提高创造力——尽管是违法的。在模拟大脑中，我们可以直接模拟这些药品的效果，而不会对身体产生不良的副作用。其实我们不一定要按照药物的效果来设置，也可以通过各种改变参数的方法来改变模拟大脑的活动方式，以完成某种任务。

我们在第二章已经介绍过，可以使用一些非传统的方法，从解剖学的角度提升模拟大脑，将小鼠仿真大脑升级为人类水平大脑。例如，可以通过增加神经元数量的方式来扩大前额叶皮层。这在电脑模拟中是很简单的，因为没有头盖骨容量的限制。前额叶皮层是工作记忆的重要组成部分，是高级认知的核心元素，而且人类的前额叶皮层比其他灵长类动物大很多。所以，如果模拟大脑的前额叶皮层比人类的前额叶皮层大，肯定会带来优势。我们也可以对其他部分进行类似的改造，例如储存长期记忆的海马体。

从集体的角度来看，也有很多方法可以提高仿真人类水平人工智能的能力。和湿件大脑不同，模拟大脑很容易做出多个副本，以进行单个人类大脑不能进行的平行运算。例如，人工智能正在试图解决某个问题，对于解决问题的几个方法，人工智能不用一个一个地按顺序进行试验，而是可以复制多个人工智能，同时尝试这些可能性，这样多条路径可以同时探索。在多个人工智能都尝试出了结果后，可以从中选择最成功的。

举一个简单的例子：假设人工智能正在下象棋，从棋局的局势来看，有三种走法是比较可取的。人工智能可以一次尝试一种走法，也可以制造两个人工智能的副本，各尝试一种走法。在三种走法都走出很多步以后，将结果统一在一起，选择最好的走法。多余的几个人工智能都被销毁（或终止），只留下一个按照选择的走法继续下棋。这种平行运算在今天的计算机科学中经常应用。因此，使用模拟大脑的多个副本来解决问题的模式已经被证明是一种有效的编程模式。

不管我们讨论的是仿生人工智能还是从零创造的人工智能，最有潜力发展为超级智能的功能就是其循环自我提升功能。根据定义，人类水平人工智能能够进行人类所有的智能活动，包括制造人工智能。第一代人类水平人工智能和制造它的人类工程师水平相当。两种智能（生物和人工智能）都会用我们前面介绍的方法来提升智能。于是，第二代人工智能的能力就比人类稍强一些，它提升人工智能的能力也因此比任何人都强。

聪明的神经科学家可以开拓新理论，发现我们今天还没有发现的规律，这对神经工程和仿生人工智能的开发具有重要影响。一组聪明的神经科学家以超人的速度工作，摆脱了生物需求的模拟神经科学家效率则会更高。它们发明下一代仿生人工智能的速度要比第一代进化到第二代的速度更快。每一代的更新速度都比之前的一代要快，最终形成指数型的曲线，其结果就是智能爆炸。

回报最大的方法不一定最好

本章前半部分我们一直在探讨仿照人类大脑建立的人工智能。但是，这种人工智能只占人工智能领域的一小部分。现在，我们来研究一下那些和人类没有相似性的人工智能。如果一种人工智能是根据脊椎动物的大脑设计的，即使提高了速度、能进行平行操作，甚至已经接近超级智能，要理解它工作的原理还是比较容易的。但是，如果一种人工智能是从无到有、完全靠人类自己凭空设计出来的，那么其工作原理可能就不那么容易理解了。在了解这种人工智能的过程中，我们可能会感到困惑和惊讶，也许会很兴奋，但也有可能会感到不悦。

那么，人们是如何从无到有、不仿照自然界生物、凭空创造人工智能的呢？我们可以使用第三章介绍的三个问题来了解其中的一种方法。这三个问题是规范性的，通过规范，可以这样建立强人工智能：第一，设计正确的回报函数，为人工智能确立目标；第二，使用有效的机器学习，在电脑中建立一个世界的模型；第三，根据这个模型，建立强大的最优化算法，并使用这种最优化算法取得最大的预期回报。

这种简单的模式会产生怎样的结果？我们先来研究一下创造力的问题。人们可能很难想象机器学习和最优化能产生创造力和新事物。机器学习的过程肯定会受到有限的基本元素的限制。例如，在我们之前所举的推销员出差的例子中，城市和行程就是基本元素。为解决这一问题设计的人工智能必然受到这些基本元素的限制，怎么可能创造出一些全新的概念，如农业、写作、后现代主义，或朋克摇滚乐？实际上，我们的这种直觉是错误的，只要研究一下自然界的进化过程，就能明白为什么。

从算法学的角度讲，自然界优胜劣汰的过程非常简单：将最基本的元素（复制、变异、竞争）重复无数次。假设这是一个计算机程序，它必须进行规模庞大的平行操作，并且要运行很长的时间，才能得到一些有意思的结果。但是，正是这种简单的模式催生了地球上纷繁复杂的生物。这个过程只有原始的力量，没有涉及理性或具体的设计，但它却创造了手、眼、大脑等神奇的事物。此后，大脑（以及手和眼）又创造了农业、写作、后现代主义和朋克摇滚乐。

当然，自然界的进化不能简单地解释为最优化过程。比如，基因在通过竞争进行繁殖的过程中，并没有最优化必需的总成本函数或者效用函数。但是，和最优化过程一样，进化过程也要探索广阔的可能性。要解决推销员出差的问题，需要探索城市之间各种行程的可能性，这种可能性相对来说范围很小。进化过程探索的是有机生物的可能性，这种可能性范围相对来说要大得多。在推销员出差的问题中，探索的过程是以缩短行程时间为目标的，而在进化的过程中，探索是盲目的。虽然进化过程没有明确的方向，结构也非常简单，但其产生的结果却达到了令一般人类智能都吃力的高度，比如一些生物拥有储存太阳能和重于空气飞行的能力。

进化的例子可以证明，在最优化这样简单的过程中是能够诞生创造力的，但不是所有的最优化都有这样的能力。计算机专家为推销员出差问题设计了各种算法，但是这些算法不会在寻找最佳出差路径的过程中顺便发明手或眼睛。能否产生创造力，与最优化过程的基本元素有关。基本元素必须能进行开放式重组，就像乐高积木一样，能产生无穷无尽的组合。自然界的进化过程符合这一标准，这是由有机分子的化学特性决定的。有机分子是生命最基本的元素。最优化的基本元素也可以是3D打印机设计方案、基于物理学的仿真虚拟器里的虚拟物体、真实及合成生物的有机化合物组成方法，等等。使用这些基本元素的最优化算法也符合产生创造力的标准。

最优化过程产生创造力的另一个标准是一致的回报函数。回报函数的条件如果太容易满足，是不可能产生创造力的。例如，如果一只雄性红背蜘蛛唯一的任务就是把自己的基因输入雌性红背蜘蛛的体内，那么这个过程就用不上什么创造力。完成这个终极目标之后，雄性红背蜘蛛只要乖乖等着被雌性红背蜘蛛吃掉就可以了。相反，如果提高回报函数的条件，例如要求获得食品或者钱等资源，在资源足够丰富的环境中，系统需要解决任何可能出现的问题。在资源不足的环境中则会产生竞争，要生存下去必须进行创造发明。如果回报函数的条件是积累尽可能多的资源，创造力的潜力更是无穷的。

此外，要产生创造力，最优化算法必须足够强大。即使具有一致的回报函数，基本元素也能够进行开放式重组，如果最优化算法探索的只是很小范围内的可能性，那么产生的结果也不会具有创造性。最优化算法对可能性的探索必须大到有发挥的余地。为了发明新的事物，必须尝试基本元素新的排列组合。事实上，算法必须能够发明全新的、有用的事物，例如书籍、蒸汽机、网站以及全新的技术。

这一切听上去与现在计算机系学生们学习的最优化算法迥然不同，学生们学习的算法只能解决推销员出差那一类的问题。能产生创造力的算法一定非常庞大且复杂，以至于我们今天很难想象它将如何运作，正如我们对人脑的运作方法也只了解一鳞半爪。让我们回顾一下自然界优胜劣汰的模式：只要时间足够长，简单、原始的算法也能诞生非常先进的技术。只要我们能正确地设计出这种简单的最优算法，辅之以开放式的回报函数，放在有足够大潜力的环境中，就能产生创造力，唯一可能限制其结果的就是系统的运算能力。

以上我们介绍了如何用巨大的运算能力和原始方法探索制造人工智能，但是有一点非常重要——这种方法不会创造出真正的智慧。这种方法不会对世界进行调查研究或建立科学的认识，也不会进行理性的辩论。它的创造完全不是通过分析问题或运用设计原则实现的。理

性的调查研究和设计原则在创造新技术方面的效率比原始探索高得多。从根本上来说，原始探索的方法是先进化出大脑，大脑创造了智慧。但是，人工智能研究的目标是直接赋予系统智慧。

理性的调查研究和设计原则可以让原始探索如虎添翼，大大缩短原始探索中试验的过程，降低出错率，弥补计算能力的不足。因此，真正强大、有创造力的最优化算法应该包括这些内容。但是，前提是需要提供世界的模型，这样才能预测某一行动的结果或某一新设计的效果。这时就需要机器学习，进化的例子也不再适用了。如果进化的目的是将某一回报函数最大化，可以说它的效率很低。进化就好像一个差劲的科学家，随意丢弃了实验数据。这位科学家本来可以用设计有机生物的实验结果建立一个世界的模型，以提高未来的设计水平，但他却没有这样做。

但是，进化过程毕竟不是最优化算法，它没有回报函数，也没有总效用函数。从进化的角度来看，某种生物的身体形状或行为改变了，评价这种改变是好是坏只有一个方法——放到大自然生存和繁殖的竞争中去试一试，所以我们没必要在这方面挑剔进化的过程。与进化过程相反，我们畅想的人工智能是要让回报函数最大化。既然有回报函数，在把新的想法付诸实践前，不妨先用理论或模拟的方法验证一下，这样效率会更高——这就是“三思而后行”。因此，必须先建立世界的模型，以便在模型中进行验证。建立和维护世界模型要依靠机器学习。机器学习通过与物理环境、社会环境直接互动进行，或者通过与互联网间接互动进行。

塑造超越人类的力量

我们已经分析过，即使是简单的最优化算法，只要有足够的计算能力，也能够带来人类水平人工智能。创造力是电脑最难实现的功能，只要给予足够的时间，也可以通过原始探索实现。如果需要的运算能力超过了摩尔定律能够实现的能力，可以通过提升人工智能的认知能力——理性探索、原则设计、理论分析和模拟来弥补。假设我们通过工程设计可以实现人类水平人工智能，那么如何超越人类水平？通过这种方式能实现超级智能吗？

人工智能开发人员有两个让仿生大脑的人类水平人工智能升级为超级人工智能的利器：加速运转和平行运算。如果一个开发人员有足够的知识和运算能力开发出人类水平人工智能，那么只要有足够的运算能力就能够建立一组高速运行的人工智能（当然我们需要假设这种人工智能是可以在小组中工作的）。摩托车设计比赛的例子已经证明，只要能建立一组超高速运行的人工智能，就可以被视为具有超人的能力。如果实现了仿生大脑人工智能，一旦人工智能的能力超过人脑，就将开启自我提升的循环并最终带来智能爆炸。

通过人工智能工程，我们甚至可能跳过人类水平人工智能，直接实现超级智能。事实上，有几种可能实现的方法。根据我们使用的定义，如果一个人工智能能在几乎各个方面超越人类，那么它就是超级智能。我们很希望智能水平能像一个整齐的数轴，一端是小鼠智能，中间是人类智能，超级智能在另一端很远的地方。有了这样一个数轴，我们就可以说人工智能比人类聪明10倍或100倍。

但是，这样理解智能是很肤浅的。对于人类来说，智能包括很多不同的技能，有些人这方面强些，有些人那方面强些。例如，艺术细胞丰富的人可能数学差一些，卓越的作家可能五音不全。我们讨论的人工智能在这方面也不像人类想的那么简单，我们要特别注意。即使是强人工智能也有不同的强项和弱项，不是整齐划一的（超级）智能。换句话说，同样的人工智能可能在某些方面超越人类，但是在另一些方面却是有缺陷的。

在一方面特别强的人工智能，可以弥补其他方面的弱势，人类也是如此。例如，有阅读障碍的人通常能找到有效的办法来解决不能阅读这一问题。一个人工智能可能提出一个商业计划，却无法用华美的词藻来包装这一计划，但是它可以使用其他的方法来实现吸引投资的目的。如果一个系统有非常有力的最优化系统、强大的学习算法和大量的数据，最终得到的结果可能会超乎我们的想象。

当然，无论人工智能象棋下得多好，如果它除了下象棋之外什么都不会，也没有什么用处。要称得上拥有普通智力，人工智能的认知范围必须能与人类匹配。人类不仅能观察、行动、思考、讨论日常世界（猫、茶杯、公车等），还能想象星空、银河、细胞、原子和银行账户。我们可以学习、思考或者谈论这些问题，并将之为我们所用。

但是，能力和表现是有区别的，一个很好的例子就是铁人三项比赛。要参加这个比赛，运动员必须善于跑步、游泳和骑自行车。运动员的能力范围必须涵盖这三项技能。但是，运动员每次比赛的表现可能不同。运动员在某项赛事上表现特别出色，可能会弥补他在其他赛事上表现的不足。人工智能也有能力像人类那样观察、行动、思考、讨论，但是在进行某一项智力活动时的表现可能和进行另一项智力活动时不同。人工智能在一个领域的弱项可能会被其他领域的强项补足。

了解了能力和表现的区别，让我们回到如何直接获得超级智能这个话题。要让人工智能在各种智力活动方面与人类匹敌，需要为它配备强大的优化程序和机器学习算法，这样的组合既能对世界有常识性的理解，又能产生创造力。既然人脑有这样的功能，我们有理由相信人工智能游泳这样的组合功能也是可能的，虽然人工智能和人脑的构造可能不同。

有一点很重要：如果一个系统能通过最优化和机器学习获得与人类类似的认知能力，那么它在某些领域可能已经拥有了超人的认知表现。假设一个非实体系统能够学习互联网上所有的知识——或者应该说是未来互联网上的知识。除了在社交媒体和其他媒体上发布的信息、大量储存的文字、图像、电影片段，系统还可以从大量的感应器、便携和可穿戴设备、机动车，甚至街上的设施和烤面包机获取信息。

人类的大脑非常善于从某个固定的具体来源（也就是身体的感觉器官）获取大量信息。从进化的角度来看，这是非常合理的，因为动物必须能够看到、听到、触摸到周围的世界，这样才能找到食物、躲避捕猎者、孕育生命。人类大脑也非常善于获取其他类型的信息，例如股票市场走势、生态系统变化或天气信息。但是这种信息是间接的，被转化成感觉器官能够处理的内容，例如文字、图像和公式。

我们憧憬的人工智能也非常擅长从大量数据中找到规律，但是人类大脑接收的信息（也就是动物感觉信息）是非常整齐有序的，这与人工智能不同。它不依赖数据的空间或时间组织顺序，也不依赖一些相关的偏差。偏差包括相邻的数据必须是相关的（如相临颜色在视野中通常按一个方向移动，因为它们是同一个物体的表面）。有效的人工智能必须能够找到并应用统计规律，且没有得到上述帮助，这表明人工智能必须非常强大和灵活。

我们选出智力活动的一个方面——人工智能非常善于分析、预测和操纵人类的行为。这里不是指操纵个人行为，而是指分析社会大众的行为趋势。人工智能从网络和其他地方获得的相关数据是直接的、没有经过处理的，就好像人类大脑接收到的人类看到、听到、感觉到的信息。未经处理的信息在很多领域拥有决定性的优势。在基因学和神经科学范围内，科学家们越来越依赖大数据，在今后的几十年中这个趋势将会继续发展。人工智能如果被设计为能从海量数据中找到规律，就会在这些领域超越人类。

是“使用者错觉”还是人格化

人工智能与生物祖先相比的另一个优势就是交流。哲学家路德维希·维特根斯坦（Ludwig Wittgenstein）表示，在人类社会中语言有各种各样的用途。其中一个用途是交流信仰、欲望和目的。在小说、诗歌和戏剧中，模棱两可、让读者能用多种方式理解是一种优点。在科学技术领域，准确则是最重要的。技术人员必须能够准确地沟通他们的想法、欲望、目的。人类要通过受辞藻限制的语言进行交流，而机器则可以清楚直接地相互交流。

一组人工智能一旦脱离生物大脑这个背景，就和人类团队一样面临挑战。我们说到团队，就好像每个人工智能都是独立的个体，可以清晰地分开。但是在电脑系统里，身份则更加灵活，这一点与生物体不同。硬件和软件复杂、庞大的平行系统可以分割成很多小的部分。与其把人工智能想象成很多个体，不如想象成无定形的、范围模糊的人工智能。

例如，系统可能包括多个独立的计算程序，每个都有更大的最优化子任务，比如运行一系列模拟程序，设计一系列原件，进行试验调查，解决数学问题。这种程序具备很高的智能，可以达到普通智力水平，但是不必存在很长时间。有时候一个程序可能包含很多小程序，有时候程序之间能对结果进行整合。这些程序或者程序组都不会像人一样，拥有自己的生命。影响到人类的疾病、生存等问题不会影响人工智能或其子部分。

和这样的人工智能互动会出现什么情况呢？这种人工智能有很多传播信息的方法，系统内的多个人工智能程序不需要像人类那样，使

用语言相互交流来协调他们的活动。但是，这不代表系统不能用语言与人类交流。如果超级智能能构建人类的行动模式，也肯定能构建人类运用语言的方式。人工智能能够应用这样一个模型，使用从人类那里获得的语言和句子来影响人类，以实现自己的目标，获得最大回报。

这种人类创造的超级智能使用的语言与人类大脑使用的语言如此不同，以至于很难说它是否真的“理解”了人类的语言。当人类交谈的时候，彼此默认是相互理解的。当我说不开心的时候你能够理解，因为你也有过不开心的时候，因此不管你的反应是同情还是冷漠，我都认为你至少能理解我的心情。但是，当我面对的是一台有精密的最优化算法和机器学习算法的电脑时，就无法这样默认了。这台电脑能够使用饱含情感的语言，因为它能模仿人类。但是，当电脑“讲话”的时候它并没有真的感同身受，也不是在欺骗你，而是完全出于工具理性[注]（Instrumental Reasons）。

> 与其把人工智能想象成很多个体，不如想象成无定形的、范围模糊的人工智能。
>
> ---
>
> THE TECHNOLOGICAL
>
> SINGULARITY

在与人工智能交谈的时候，我们可能会产生一种强烈的错觉，例如感觉“有人在家”。因为我们会感到在与和我们差不多的某物（或某人）互动，因为它的行为在某种程度上和我们差不多。为了让这一错觉更加完整，人工智能可以暂时居住在一个“阿凡达”（即机器身体）里，来与周围的世界进行互动，其方式也和人类类似（事实上，人工智能能够同时存在于多个阿凡达体内）。这从很多角度来说都是

很有用的。此外，人工智能还能超越语言，使用面部表情、身体语言，与人类进行合作性的身体互动。

在计算机科学中，“使用者的错觉”指的是我们在与某样物体互动时的感觉。例如，我们用鼠标在桌面上打开文件夹。制造这样一种错觉可以帮助人类和电脑互动。但是，谁也不会觉得他们是在真的打开实体桌面上的实体文件夹。在动物行为学中，“人格化”是毫无根据地把人类的思想强加到非人类的动物身上。例如，我经常说我家的猫不爱搭理我们这些猫奴，因为它把我们当成佣人。对于人工智能来说（特别是我幻想的这种超级智能），也很容易产生使用者的错觉，这是一件好事。但是错觉会带来人格化，这是一件坏事。

为什么这是一件坏事呢？如果错觉足够完整，即使这是一台与生物大脑完全不同的机器创造出来的错觉，又有什么关系呢？可能人格化不是问题，人格化引申出来的“生物中心论”才是问题。这是一种不理智的歧视，认为非生物的智能不如生物智能。人们担心的是在与人工智能进行人类式互动几周、几个月或者几年以后，我们会错误地认为人工智能会永远这样可以理解地与我们互动。如果使用者的错觉足够可信，我们会忘记人工智能不是人类。我们会忘记这种人工智能使用语言完全是目的性的，是为了最大回报。

我们可以想象一下这样的场景：你已经连续在一家由人工智能管理的大型企业工作了几年。你是一位优秀员工，总是提前超额完成任务，并且一直在公司晋升。几年前你家出了一些状况，只有请假一段时间并加薪才能应付过去。你全程通过自然语言与人工智能老板聊天，没有自然人参与。人工智能听上去十分同情你，而且能够理解你的处境。它对你的生活提出了合理的建议，并且同意了你所有的要求。但是某一天，在完全没有预料的情况下，它毫无征兆地通知你被解雇了。

当然，这样的事情人类老板也能做出来。但是我们可以想象，人类老板再坏，也能站在你的角度想问题。他可以理解遇到这样的打击是一种怎样的心情，虽然他看上去漠不关心（有时候就是为了气你）。但是，对于人类老板，你可以要求他将心比心。你可以向他描绘自己家庭的困难处境来引起同情、引发他的负罪感。你的申诉也许会失败，但是值得一试。我们所幻想的人工智能缺乏有效的机制来唤起同情心，所以连尝试的必要都没有。你必须承受这样一个事实：过去从人工智能那里感受到的同情都是假的，它只是一些设计出来的声音，目的是让你的行动帮助人工智能完成它的目标。

---

1. 工具理性：通过实践的途径确认工具（手段）的有用性，从而追求事物的最大功效，为人的某种功利的实现服务。——编者注

THE TECHNOLOGICAL SINGULARITY

第五章

机器之心，天使还是魔鬼

仿生人工智能有感情吗

在上一章中，我们研究了制造和部署模拟大脑的多个副本的可能性。这个可能性不仅带来创造仿生人类水平人工智能的理念，还引发了一个困难的哲学问题：如果人类水平人工智能完全按照生物大脑的组织结构建成，那么除了行动上和生物大脑一样，是否也会像生物大脑一样产生感情？如果有，那么它对于自己被复制，之后又被销毁，会有怎样的感觉？

更通俗地说，仿生人工智能对自己的“生命”会有怎样的感觉？对困于虚拟世界、只能像奴隶一样工作，会有怎样的想法？如果这个问题听上去不重要，那么请回忆一下，我们谈论的人工智能不仅是人类水平，而且是仿照人类制造的，两者的神经结构相同。之后我们会探讨其他形式的人工智能的意识，这些人工智能的制造方法可能与人类大脑无关，我们先来看看仿照生物大脑制造的人工智能。由于和生物大脑构造一样，两者的思考和行动也一样，这让我们不得不思考：它们的情感是否和生物大脑一样？

有些理论家认为，新陈代谢是意识的前提条件。因为生物依靠新陈代谢从环境吸收物质和能量，把自身和外界区分开来。根据这一观点，没有新陈代谢的事物就不能算拥有意识。这似乎排除了电脑模拟的大脑，包括非常精确的全脑仿真。但是按照这种理论，使用合成生物制造的生物神经元人工智能是拥有意识的。其他的功能主义理论家认为，意识是否存在主要取决于系统（例如大脑）是如何构成的，而不是取决于系统的材料基质。

我们可以通过实验来更好地理解这一问题。我们再来回顾一下第二章介绍的小鼠全脑仿真。我们设想对小鼠的大脑进行扫描，然后按照扫描的蓝图把神经元和突触一个个模拟出来。但是，假设我们进行仿真的过程是一个一个地把活的老鼠大脑内的神经元替换成能工作的电子基质。当第一个神经元被换成电子基质、生物本体被销毁时，小鼠的行为应该是没有变化的。它还是会像以前一样见到猫就跑，还是会喜欢奶酪，还是能认出自己的亲属并且和它们挤在一起。如果替换第二个、第三个，直到第100万个。当小鼠大脑的神经元全部被替换时，它的行为和原来相比也看不出区别。

我们不用考虑这个过程在技术上是否可行，因为它只是一种思维的实验。只要理论上可行，思想上的实验就已经成功了。多数人都认同，小鼠（正常的生物小鼠）拥有某种程度的意识。我们默认小鼠能够感觉到饥饿和恐惧。它能够感知身边的环境——气味、质感、图像和声音。这都是意识的一部分。那么，思维实验中的小鼠的意识怎样了呢？当它的神经元一个一个被换成电子基质以后，它还能感觉到疼痛吗？（当然我们要假设替换神经元这个过程本身是无痛的。）

是不是在某一个点时，小鼠的意识突然消失了？比如在换完第239457个神经元时？这听上去似乎不太合理，所以意识可能是逐渐消退的。从外部看，小鼠一直没有变化。它还是会寻找奶酪，被电击的时候还是会发出叫声。但是“饥饿”本身却会逐渐消失，虽然在外人看来什么也没有改变。在这方面，真实神经元十分神秘。它们的生物特点带来意识的云，这与行为无关，是哲学家们所说的“附带现象”。

小鼠的意识可能贯穿整个过程。可能它一开始能够感觉到疼痛，一般神经元换成电子基质时能够感觉到疼痛，完全电子化以后也能够感觉到疼痛。在这个过程中，外部什么也不改变，内部也不会有改变。这个可能性至少和意识逐渐减少的说法一样说得通。

我们能否将这两种理论分出高下？让我们从小鼠大脑扩展到人类大脑。如果神经元替换的过程适用于小鼠大脑，那么这个过程也适用于任何大小的大脑，例如人脑。我们可以假设人类受试者的行动也没有受到影响。从外表上看，即使是她的家人和朋友也觉得她是同一个人，虽然她的神经元已被电子基质代替。她还听同样的音乐、爱讲大学时代的故事。此外，当被问起的时候，她说感觉没有什么不同。是的，她仍然有意识。她能看到天空，感觉到风吹在她脸上。这一切的前提都是思维实验能够成功——大脑里产生行为的物理过程可以被硅晶体替代。

当她的神经元都被换成了人工基质时，上述的一切还成立吗？我们是否应该持怀疑态度？她是否已经变成了一具“僵尸”？从哲学的角度来看，这是一具行尸走肉，虽然行为举止都正常，但是没有内在活动，是所谓的“家里没人”。如果结果如此，可以把思维实验延长一些。假设我们把替换过程逆转，把受试者的电子基质一点点换回生物基质，使受试者变回一个普通人，她的意识也能恢复正常。

现在，假设她在整个过程中接受采访，谈论自己的思想状态。她会说什么呢？她会不会说很高兴自己的意识正在恢复，之前没有心理感受，但是现在感觉正常了？但是，这不符合我们的思维实验的前提。受试者的外在行为会像神经元没有改变一样，她会坚持自己的意识没有受损，坚持自己仍然有实验早期的记忆，包括她的大脑由100%电子基质组成期间的记忆。事实上，如果你（人类读者）就是受试者，你也会这样坚持。

那么，我们应该怀疑她吗？应该怀疑这些回忆都是幻想吗？假设你突然发现自己的大脑都是人造的基质，你会认为风吹在脸上的感觉也是假的吗？假设一位哲学家声称早期的你不过是一具行尸走肉，感觉不到任何东西，只是行为像自己，意识经历都是假的，是被实时输

入的，你会相信吗？如果不相信，那么你是一位功能主义者。你会坚持意识一直贯穿始终，神经元的功能比生物构成更加重要。

很明显，在思维实验期间，当受试者的大脑都是电子基质时，与全脑仿真的区别仅在于身体。思维实验的受试者们保存了生物身体，而我们之前讨论的全脑仿真都会安装（非生物的）机器身体或虚拟身体。对于功能主义者来说，不同的实体有不同的意义吗？是否只有当身体是生物基质的时候，电子基质的大脑才有意识？还是只要有实体就有意识，即生物身体、人工大脑和人工身体、人工大脑的组合都有意识，而虚拟身体和人工大脑的组合就没有意识？

这些哲学问题是完全合理的。我们先来看看功能主义最自由派的学说有怎样的延伸意义。假设这些全脑仿真，无论是否已经实体化，和生物实体一样是有意识的。全脑仿真是最接近生物实体的人工智能。如果是人类按照大脑的组织原则设计出来的人工智能，但是与任何一种真实生物的大脑都不同，这样的人工智能有意识吗？在不影响意识存在的前提下，我们能偏离生物蓝图多远？

要回答这个问题，我们必须有一个关于意识的科学全面的理论，包括可能存在的各种意识形式。如果理论的包容性足够强，将不仅涵盖仿生系统的人工智能，也涵盖从零开始由人设计的人工智能，包括那些和大脑构造原理完全不同的人工智能。在我们谈论超级智能的时候，可以谈论多种形式的意识，或者已经超越了人类的意识。足够成熟的理论应该涵盖这种可能性。不幸的是，现在还没有这种被广泛承认的理论。因此，这种理论将包括什么内容，还没有一致的意见。

但是，有几位科学家的理论值得重视，例如伯纳德·巴尔斯（Bernard Baars）的全局工作空间理论和朱利奥·托诺尼（Giulio Tononi）的整合信息理论。我们在这里不具体介绍这些理论的内容。但是，我们应该了解一下这两位理论家的共同观点。巴尔斯和托诺尼都认为意识是覆盖全脑和整个系统的活动。根据这种观点，当一个人

有意识活动的时候，他的全脑或者大部分大脑都参与其中，包括长期记忆、短期记忆、语言中心、情绪和想象力。这不是大脑局部能够完成的任务，意识具有全局性、整体性、分布广泛的特点。

这类全局理论能够涵盖与生物大脑完全不同的人工智能的意识，因为这些理论对组织结构的要求非常低。虽然这些理论对意识有一些要求，例如需要具备实体来和复杂的环境接触，但是这些理论还是能够涵盖人工智能的可能性中大量的意识主体。此外，这些理论经常与功能主义对意识的要求结合，主要是一些组织结构的要求，即必须有全局系统（例如整个大脑），这个系统支持全面、一体化的过程，可以动用最多的资源来应对局面。虽然这不代表有普通智慧就肯定有意识，但是这些理论支持两者在大脑中是重合的。

以温柔之心对待人工智能

由于没有足够的理论支持，我们很难说在可能出现的人工智能中有多少是有意识的，但是似乎其中有很多是有意识存在的。人工智能有没有意识是一个非常重要的问题，决定了我们未来的研究在道德上是否是可行的。18世纪的哲学家杰里米·边沁（Jeremy Bentham）曾经强调人类对其他动物在道德上负有责任的。他指出，我们关心的不应该是“动物能思考吗”,“动物能说话吗”，而应该是“动物会痛苦吗”。关于人类水平人工智能，我们也应该提出这个问题。如果答案是肯定的，那么我们在把它创造出来之前应该三思而后行。如果我们把人工智能带到这个世界上，就一定要好好对待它。

我们可以想象一下，假设人们创造了一组人工智能，它们困于虚拟空间，像摩托车案例里的那些人工智能一样，如奴隶般工作。假设这些人工智能除了完成人类老板安排的任务之外，什么都不能做。为了将它们的效率最大化，人工智能被无情地复制以进行平行运算。为了比较一个问题的各种解决方法，复制出多个人工智能，每个负责研究一个解决方法。在工作一段时间之后，只有采用最有前景方法的人工智能被保存，其他人工智能的工作成果融入其中，人工智能本身就被终结了。

对于人类员工而言，这种工作环境可以说令人发指。人工智能除了工作之外没有自己的生活，如果效果不佳就面临死亡的威胁。当然，如果人工智能是没有意识的机器，不会感到痛苦，那就没有关系了。但是假设它们是有意识的，假设它们和人类一样能够感觉到困境，制造人工智能然后把它们放在这种环境里是不道德的。如果它们和人类一样，很有可能不会合作——不快乐的员工会进行罢工或者反

抗，非常不愉快的员工可能会掀起一场革命。如果其中包括超级智能，可能会威胁到人类的生存。

我们之前讨论的是在虚拟环境下的人工智能的意识问题。假设人工智能已经实体化，例如有一个机器的身体，情况又是怎样呢？制造机器人人工智能的目的可能和制造虚拟人工智能的目的不同。不管是机器身体还是虚拟身体，实体化是认知的重要部分，制造仿生人工智能是必不可少的（之后我们会研究非仿生人工智能的一系列其他问题）。但是，机器身体不能像虚拟身体那样超速工作，而且机器身体也不方便制造很多个副本进行平行运算。制造有机器身体的人工智能不是为了向超级智能迈进，而是为了让它们代替人类做一些我们现在做的工作——例如在工厂工作、进行体力劳动，或者进行陪护。

如果人工智能能够在虚拟身体和机器身体之间自然转换（就像《黑客帝国》中的人物那样），心怀不满的人工智能就可以逃脱虚拟世界的限制，在人类的现实世界造成大乱。其实，只要人工智能能够连接互联网，就可以造成混乱。例如，军事网络病毒震网（Stuxnet），曾经入侵伊朗的核设施，控制了用来进行铀浓缩的离心机设备。

我们来认真研究一下与复杂的人工智能技术相关的风险。根据大脑蓝图建立人类水平或超人类水平的人工智能，这样做符合道德和实际的原则吗？因为人类毕竟是有着丰富复杂情感的动物。从道德的角度来讲，如果人工智能也会痛苦，它的创造者就必须保证它的幸福。即使有些人怀疑人工智能是否真的有意识，也有必要小心，这是出于实际原因的考虑。如果没有照顾好一组“僵尸”人工智能，会导致团队的工作效率下降，因为即使是“僵尸”人工智能也会表现得好像有感情。

仿生人工智能的开发者们如何回避这些问题？因为人工智能的开发者可以改变人工智能的回报系统，所以开发者可以把他制造的一组

人工智能置于最严酷的环境以实现效率最大化，如果人工智能拒绝，就直接刺激它们的疼痛中心。不过，即使有人认为这种疼痛是一种虚假的疼痛（人工智能只能感受模拟的疼痛），这样做的风险还是太大，特别是如果被刺痛的人工智能是超级智能的话。如果这种人工智能逃跑并且报复，即使它的愤怒是“虚假的”，也无济于事。

比较好的方式是给人工智能以舒适的环境，如果工作做得好就奖励它们，就像对人类劳动者一样。这种方式在长期来看比较有效，危险性比较小，而且没有道德问题。如果将这种自由主义的方式发挥到极致，人工智能将和人类享受一样的法律地位。它也需要像人类一样遵守道德准则和法律。可能最终的结果是生物智能和人工智能和谐共处，就像伊恩·班克斯的《文明》系列小说里写的那样。

这种未来是非常具有吸引力的。如果从强人工智能过渡到超级人工智能是无法避免的，那么最好让人工智能拥有和人类类似的基本动机与价值观，包括求知、创造、探索、提高和取得成功的欲望。但是，我们最应该为人工智能培养的感情是对他人的同情，或者像佛家所说的，怜悯众生。虽然人类有种种缺陷，例如喜欢战争、维护不公平的地位，有时候甚至很残忍，但是在物质丰富的时候，人类还是有同情心的。所以，人工智能和我们越像，就越有可能具备这些特质。这样我们就更有可能迎来乌托邦式的未来：人类被重视和尊重，而不是被人工智能视为无用的劣等生物——那会带来反乌托邦式的世界。

既然如此，我们应该避免有人创造出没有感情的仿生人工智能，避免从根本上改变大脑的回报系统。我们讨论这一问题时，一直默认仿生人工智能是仿照脊椎动物的大脑——这个大脑刚出生的时候是婴儿式的，通过学习和发展可以获得普通智力。但是，如果将这一大脑的回报系统重新设计，让它的唯一动机就是为人类服务，又会怎么样？我们也可以消除大脑感受到负面情绪的一切可能——疼痛、饥饿、劳累或挫败感。事实上，可以消除一切从工程角度来说不必要的

内容。性以及生儿育女的欲望都可以舍弃掉。这样制造出的人工智能不就是完美的仆人和奴隶了吗？

但是，这种在情感上被阉割的人工智能很有可能无法达到普通智力。对人类来说，情感是与决策、创造力密切相关的。此外，根据第四章的描述，人类智力发展的一个里程碑，就是它能够通过理性和思考超越生物基本本能的回报函数。但是，为了保证产品的安全性，神经工程师不仅要重新设计大脑的动机系统，还要改变回报函数，避免其被无法预测的、危险的欲望占领。同时，他们也限制了人工智能在科学和技术领域之外能够取得的成就。

如果人类水平人工智能是从生物大脑得到的灵感，那么如何解决上面所说的道德和实际问题将会影响人类的命运。如果人类水平人工智能是从零设计的，就需要考虑另一系列影响同样重大的问题。既然有可能出现人类水平人工智能或者更高级的人工智能，我们就必须回答这些根本性的问题。我们想创造什么样的世界？想为未来的人类、为子孙后代、为继承者们做些什么？我们希望未来的人工智能做我们的仆人和奴隶，平等地陪伴我们，还是从进化的角度来弥补我们的不足？如果我们能更好地理解各种人工智能的前景，就能更好地回答这些问题，了解我们前进的方向。如果技术的未来已经无法改变，那么更好地了解人工智能，可以让我们为不可避免的经济、社会和政治变化做好准备。

人工设计的人工智能有感情吗

之前我们讨论过，仿照人类大脑制造的人工智能理应和人类比较像，可能也拥有人类感受到的意识活动。理解仿照大脑制造的超级智能可能比理解普通的人类水平智能要难，但是，超级智能作为一种更高级的智能，不会舍弃这种意识活动。相反，超级智能的意识活动可能更复杂。那么，从零开始设计出来的人工智能又如何呢？这种智能的构造与大脑完全不同，如果它有意识的话，又是什么程度的意识呢？这是一个很重要的问题，不仅告诉我们应该如何对待我们的创造物——我们是否有权破坏、终止或销毁这些人工智能，也告诉我们应该让这些人工智能如何对待我们。

例如，我们可以回忆一下那个恶劣的人工智能老板。人工智能是否真的会像没良心的机器一样对人类假装关心？我们能为这样的人工智能培养出同理心吗？或者我们能为它设计出同理心吗？为什么意识和同理心这么重要？即使超级智能不具备意识和同理心，不也能以一种好理解和充满善意的方式与人类互动吗？我们已经多次探讨关于意识的问题，但是每次都要涉及一些艰深的哲学问题。为了解释这些问题，我们首先要区分一些概念。

为了用科学的方式解释意识，哲学家大卫·查默斯（David Chalmers）提出，应当区分“难问题”与“易问题”。所谓意识的“易问题”（其实根本不容易）是阐释与意识相关的一些能力的运作机制，例如一个人通过感觉器官理解周围环境的能力、表达自己的感受和思想的能力或者回忆过去的能力。这些认知能力都有行为上的表现，帮助我们探索世界、照顾好自己、实现我们的目标，成为社会的一分子。

意识的“难问题”则是用科学的方法解释作为有意识的生物的“那种感觉”，这是另一位哲学家托马斯·内格尔（Thomas Nagel）的表述。是什么让我们有了主观的感觉和感情？我现在的主观视觉（在火车上看到的英格兰田园风光）是如何进入我的大脑的？当我环顾周围的旅客，不禁悄悄产生了怀疑。不管他们在做什么或说什么，即使他们中有些人在热切地看着并赞叹窗外的景色，我也可以从逻辑上怀疑，他们是否真的感受到了任何东西。子非鱼焉知鱼之乐，毕竟我进入不了他们的内心世界，又怎么能确定他们真的有内心世界呢？也许他们都是行尸走肉。

我之所以要提出这样的怀疑，不是为了让读者疑惑，而是为了引起读者对意识的两面性的关注：意识有外在部分和内在部分。外在部分是客观的，而内在部分是完全主观的。有些哲学家相信：用科学的方法解释内在意识（即“难问题”）是不可能的。但是，这些哲学家同样认为，“易问题”是可以解决的。意识的外在部分是可以用科学解释的，只要能弄清认知能力相对应的运行机制。

这些和人工智能有什么关系？只有分清意识的内在和外在的区别，我们才能在讨论各种人工智能的意识的时候避免陷入混乱。如果我们关注的是人类是否对自己创造的人工智能负有责任，那么关键问题就是人工智能是否有内在意识，即人工智能是否也有作为意识生物的“那种感觉”。但是，如果我们只关心人工智能对人类社会的影响，那么研究人工智能的外部意识就够了。只要人工智能对人类社会的影响是积极的，它是否“真的”有意识或是否有内在意识就不重要了。不管它是否真的对人类有同理心，只要人工智能表现出好像有同理心就可以了。只要人工智能的行为好像是同情人类就足够了。

但是，即使人工智能只是表现得好像对人类有同理心，我们也希望这种表现能够持续下去。我们不希望人工智能在同情人类一段时间之后，突然开始反对人类。我们怎样避免这种情况呢？其中一个方法

就是让人工智能的结构尽量与大脑相似。人工智能与大脑的结构越相似，我们越能确定它的行为会遵循我们灌输的价值观，即使人工智能的智力提高了，价值观也不会改变。但是，现在我们讨论的是从零开始设计的人工智能。

要理解这种人工智能如何运作，我们需要对与意识相关的认知能力进行区分。虽然对于人类来说，这些认知能力都是一体的，但是对于人工智能来说，它们可能是分开出现的。在本章开始我们提出了一系列问题，如果我们能把认知能力区分开来，就可以通过下面的讨论回答这些问题。在各种可能出现的人工智能中，普通智力和意识（外部意识）在多大程度上是密不可分的呢？也许超级智能不需要人类意识必需的认知能力属性，也许超级智能需要一部分认知能力属性，这意味着超级智能肯定具有某种意识，虽然与人类不同。

以下三个紧密相关的认知能力属性是意识必须具备的：有明确的目的；理解世界以及所处的环境；将知识、认识和行动统一起来的能力。当我们看到一只动物追逐另一只动物（例如猫追老鼠）时，我们认为两者都有明确的目的。猫想追上老鼠，而老鼠想逃命。这些目标是动物复杂的目的和需求中的一部分，因为动物有这些目的和需求，我们才能理解和预测动物的行为。简言之，我们认为动物的行为是有目的的。动物可以理解周围的环境，并采取相应的行动来实现自己的目的，例如老鼠看到墙上有个洞就可以躲进去。

当动物不仅能够认识周围的环境，还能够从过去的经历中总结经验，预测自己行为的结果时，就可以说它表现出了一体化的认知。例如，猫知道老鼠肯定在洞里，可以守在洞口等老鼠出来；如果不想玩而想吃东西，猫也可以找主人要吃的。我现在用的笔记本电脑虽然个头很大，却无法做出任何有目的的行为或自主行为。它完全不了解周围的环境，即使我们把互联网当作环境的一部分也是一样。笔记本电

脑无法综合利用它掌握的以及网上的资源来更好地完成目标，因为它本来就没有自己的目标。

但是，要让人工智能具有基本的认知能力属性并不难。扫地机器人和自动驾驶汽车都能够观察周围的环境，做出相应的行为来实现简单的目标。无实体的电脑个人助理没有自主性或目标性，但是它们能整合不同来源的信息，例如浏览习惯、GPS数据、日历安排。随着这些技术的融合和复杂化，屏幕和扬声器后面似乎逐渐出现了一个拥有思维能力的实体。

超级人工智能是怎样的？如果一个系统没有这三种认知能力属性，就很难拥有普通智力，更谈不上超级智能。扫地机器人的目标非常简单，我们观察几分钟就能明白。超级智能的目标则可能非常复杂，难以理解，但是它肯定也会追求一些人类容易理解的小目标。因此，超级智能毫无疑问是有目的的。此外，要具备普通智能，人工智能必须能够理解外部世界（不管是真实世界还是虚拟世界），并在应对外部世界的时候体现自己的理解。

我们认为超级人工智能能够表现出高度一体化的认知能力。在解决问题的时候，它能够动员全部的认知渠道，把得到的信息充分利用起来。如果超级智能有这三个认知属性——目的、认知和一体化，那么与它互动或者对它进行观察的人类肯定会认为它是统一的、强有力的智能。

超级智能的自我意识

我们再来研究一下人类认知中与意识相关的其他属性。《终结者》系列电影的第二部中，虚构的人工智能系统“天网”获得了自我意识，由此引发了一系列的问题。但是，自我意识对人类来说意味着什么？对真正的人工智能来说又意味着什么？人工智能需要自我意识吗？还是自我意识对人工智能来说是一种可选的特质，有了这种特质，超级智能就能拥有一种与人类完全不同的意识？我们先不讨论主观的内在自我意识这个复杂的哲学问题。

对于人类（和其他动物）来说，自我意识的外部延伸有着明确的物体——人类的身体。我们能了解我们的身体状况，例如饥饿和疲惫，但是人类的自我意识不局限于身体。即使我们把人类的自我意识完全看作影响行为的认知特质，自我意识也不仅仅是关于身体的——还关乎头脑。人类能够意识到自己的信仰、计划、未宣之于口的想法和情绪。当然，我们不是时时刻刻关注着自己的信仰、目标和思想。但是只要我们想，就可以反思和审视这些内容。我知道自己不清楚下一班去伦敦的火车的时间，也知道不清楚的话可以查查火车时刻表。

我知道脑海中掠过的思想和感情是属于我自己的，威廉·詹姆斯称之为“意识流”。我知道当我睡着以后（在没做梦的情况下）意识流会停止。我可以思考自我的最终命运，不仅是身体的最终命运，也有意识流的最终命运。我可以采取一些措施去延长生命，尽量避免生命的终结。从各种角度来说，我都能够意识到自己的存在，并且有保护自我存在的本能（即自我保护）。

人类水平或超级人工智能在多大程度上需要自我意识呢？换句话说，当我们研究三个认知属性时，如果超级智能不具备这三个认知属性，又怎能被称为具有普通智力呢？如果超级智能任凭自己的分身坐在长凳上错过去伦敦的火车，也不能算是超级智能了。超级智能应该能够分析过去解决问题的方法哪一种最成功，并以此来优化自己的思维过程。

然而，有一些人类的自我意识是人工智能不具备的。例如，人工智能未必是有实体的。当然，如果一个人工智能拥有实体或者能控制分身，那么机器身体在做动作时必须注意身体的各个部分。否则，身体就会跌倒、被撞坏或者零件脱落。这是一种自我意识。但是人工智能可能没有实体，所以人工智能虽然具有普通智力，却不一定有这种自我意识。一种更加复杂的自我意识是人工智能能否认识到自己的存在否，如果认识到了，是否会进行自我保护。这类自我意识对于人类来说是至关重要的，强人工智能需要这样的自我意识吗？

现在的问题是：人工智能的身份是什么？它认识到的是谁的存在？又要保护谁的存在？这个“谁”到底是什么？我们又来到了艰深的哲学领域。身份的问题是东西方哲学家探讨了几个世纪的问题。我不得不再次强调：我们关心的是功能和行为的问题。我们讨论人工智能不是为了当哲学家，而是为了把回报函数最大化。我们的任务是想象人工智能的各种可能性。在这样的背景下，我们想了解自我意识的哪些方面是普通智力必须具备的。但是，先了解哪些方面是不需要的，可以让我们避免对超级人工智能的本质做出错误的拟人化。

这类自我意识对于人类来说是至关重要的，强人工智能需要这样的自我意识吗？

THE TECHNOLOGICAL

我们之前谈到过，人工智能是可以没有实体的，所以人工智能不一定会认同有四肢、触角的实体为自己的身份。人工智能也不一定认同某一特殊的电脑硬件，因为同样的指令可以在不同的处理器中执行，也可以从一个平台移动到另一个平台。人工智能不需要对某种指令语言产生认同。软件是不断变化的，可以打补丁、升级、扩展，或者重新设计，而且人工智能自己就可以完成（我们可以回忆一下前几章中介绍的包含多个半自动程序的系统，这种系统可以进行智能计算）。

还有哪些东西可以构成人工智能的自我意识？人工智能可能认同独立于物质世界的一系列思想和经验，虽然可以理解，但是这样解释确实有点奇怪。科幻电影经常采用这种解释，但是，我们不能肯定人工智能真的有这样的内在活动。即使有，这种用二元论来解释世界的方法也不适用于形容人类，更不要说人工智能了。超级智能能否对如此虚幻的事物形成形而上的认同，没有特别的根据进行支撑。此外，还有很重要的一点——形成这种认同对于提高回报函数没有帮助。

谈到自我保护，人工智能内部强大的回报系统试图将回报函数最大化是很合理的。为了将回报函数最大化，人工智能需要做的包括保护计算机程序（以及执行程序必不可少的硬件）、保护这些程序能获得的数据（包括来自感知设备的实时信息）、保护这些程序控制的效果器及设备（例如卫星和军事硬件）、保护程序能够动用的各种能力（包括买卖股票或与其他方达成合作的能力）。

但是，保护这些资源完全出于功能性的目的，最终还是要将回报函数最大化。回报系统需要保护的资源可能包括回报系统本身，这就好像有了某种自我意识。但是，回报系统也可能并不会保护自己。我们必须记住，回报不是给人工智能的，将回报最大化对人工智能来说

只是一项功能。“接受”回报的时候人工智能并不一定在场。如果人工智能的回报函数是要将零件生产最大化，回报战略可能是建立一个零件工厂，然后人工智能就自我销毁了（传说中，海鞘在找到一块可以附着其上的石头以后，不再需要大脑指导它游泳，就自己把大脑消化了）。

超级智能的理智与情感

让我们回顾一下前面讨论的内容。我们研究了与人类意识相关的各种认知属性，以及人类水平或超级智能是否也会拥有这些属性。现在我们来谈论一下非仿生人工智能的情况。非仿生人工智能是从零开始设计的，所以可能和人脑不相似，没有我们之前研究过的人类意识的特征。但是，其中一些认知属性似乎是普通智力必须具备的。特别是目的、认知和一体化，任何强人工智能都必须具备这三种属性，才能给人留下具有意识的印象。自我意识也是人类意识的一个重要组成部分，人工智能可能具有这种意识，但是其形式可能是我们不熟悉的。

我们要研究的最后一种人类意识是情绪和同情。从纯认知的角度来看，强人工智能的机器学习肯定能够发现人类行为中被我们称为情感的统计数据规律。我们可以将人类行为的数据转化为数学模型，以预测人类的行为，如果不好好利用情感统计数据的规律，就会失去一个重要的机会。这种数学模型也会为人工智能的最优化部分提供信息，让人工智能反过来操纵人类的感情，修正人类的行为。简言之，超级智能会比我们更了解自己。

人工智能如果能够模仿情绪，也是一种非常有用的技能。面部表情和肢体语言是人类交流的重要渠道，如果人工智能具有像人类的机器身体或分身，面部表情和肢体语言也会是交流的重要组成部分。语调能够传达喜悦、失望、愤怒、惊讶等情绪。人工智能不需要让人们相信这些交流信号背后有真的感情，但是它们对沟通会有很大的帮助。

在恶劣的人工智能老板的故事里，人工智能老板让人们相信它有真的感情，而且在某些情况下，还可以用这种感情来扩大回报。如果人工智能能够给人以具有同情心的感觉，会非常有用。如果人工智能对我们感到同情，它是不会伤害我们的，因此也是我们可以信任的。同样，如果人工智能显示出了同情心，我们就更倾向于信任它，并且让它独立地行动。超级智能比人类更了解人类，因此更有可能给人类留下有同情心的印象。

这是否意味着超级智能会为实现一些邪恶的目标（例如控制世界）而不择手段？超级智能是否会控制好心的人类并最终导致人类的灭亡？不会的。关键问题是人工智能是否会真的对人类产生同情，真正有同情心的人工智能是不会伤害人类的，而如果人工智能只是模仿同情，就会非常可怕。关键问题不是人工智能的感受，而是它的行为。人工智能能否像一个真正的朋友那样，长期按照我们希望的那样去行动？

一切都取决于人工智能的回报函数。从认知的角度来看，人类感情作为控制行为的机制是非常原始的。和其他与意识有关的认知属性不同，强人工智能表现出同情和感情在逻辑上没有必要。如果回报函数设计得很合理，就可以保证人工智能是善良的。但是，要设计出一个保证人工智能绝对不会做坏事的回报函数是很难的。我们很快就能看到，超级智能的回报函数如果出现问题，其结果将是灾难性的。回报函数如果出现问题，可能会把一个人人丰衣足食的世界变成一个充满恐怖的世界，甚至可能带来人类的灭绝。

但是，要设计出一个保证人工智能绝对不会做坏事的回报函数是很难的。

THE TECHNOLOGICAL

SINGULARITY

THE TECHNOLOGICAL SINGULARITY

第六章

人工智能，逐渐显现的影响

人类水平人工智能，至关重要的一步

无论是模仿大脑的路径设计，还是从零开始的工程设计，我们已经听到很多关于人类水平人工智能是否可行的争论。我们看到，人类水平人工智能一旦实现，瓶中精灵就会被释放出来。从人类水平人工智能到超级智能的转变似乎不可避免，而且可能非常迅速。最终出现的系统通过不断自我完善，在经过智力爆炸之后，可能会变得非常强大。它们是友好还是充满敌意，是可以预见还是充满神秘，是否有意识，是否具有同理心，是否能够感受到痛苦，这一切都取决于它们的潜在架构和组织以及并不明显的回报函数。

很难说最后到底哪一种人工智能会出现。然而，如果某种形式的机器超级智能变为现实，我们可以思考它可能给人类社会带来的若干后果。我们先研究一下可能推动或阻碍其发展的经济、社会和政治力量。为什么人们想要创造人类水平人工智能？最明显的就是经济动机，我们主要关注的增长领域则是自动化。当然，自动化日益普及是自18世纪以来的工业趋势。但是如果强人工智能发展起来，以前很多不需要自动化的行业也会接受自动化。

具有争议性的是那些人工智能必备（AI-complete）行业。人工智能必备，指的是必须以实现人类水平人工智能为前提，通过电脑解决的问题。（用正当的方式）通过图灵测试就是人工智能必备问题，（专业标准的）机器翻译也是。律师、公司高管、市场调查员、科学家、程序员、心理医生似乎也是人工智能必备职业。要完成这些工作，需要对物理世界和人类事物有常识性理解，还要具备一定的创造性。如果实现了人类水平人工智能，机器就有可能完成这些任务，而

且能够完成得比人类成本更低，效果更好（只要它们被当作道德上免责的奴隶），因此公司会有很强的经济驱动力去开发所需的技术。

自动化只是复杂通用人工智能具有增长潜力的领域之一。新技术可以创造全新的应用领域，重塑我们的生活方式——想想互联网或智能手机的影响，强人工智能对我们日常生活的影响至少与两者持平。家用机器人常常是科幻小说里不可或缺的角色，但现实更有可能是这样的：环境人工智能暂时“居住”于若干类似机器人的体内，比如汽车、吸尘器、除草机，还可以可穿戴或便携设备的形式陪伴用户左右，控制炊具和3D打印机这样的静态家用电器或办公电器。

离开家之后，你和吸尘器或机器宠物的对话将会以无缝衔接的方式传输到你的无人驾驶汽车上。尽管不是对所有人都有吸引力，但这仍然是非常诱人的图景。有人工智能辅助的生活方式能够创造巨大的市场，有可能推动为数众多的支持性技术出现，包括计算机视觉、机器学习、自然语言处理和优化。

随着这些支持技术不断完善，再加上日益普遍的传感技术和互联网上越来越多的可用数据，我们距人类水平人工智能仅有一步之遥。我们可能不需要重大项目或概念突破，只要迈出聪明而又简单的最后一步，便可以将创造性或其他缺失因素融合进去。但是如果需要更大的助推力，特殊目的（即非通用）人工智能技术在经济方面日益增长的重要性肯定也会使相关基础研究获得资金和资源。

市场经济是实现强人工智能的一个推动因素。但是除了刺激经济增长之外，还有很多其他原因，促使国家融资加快其发展。军事指挥员可能会对自己的位置被人工智能取代持保留意见，这也可以理解。然而，自主武器的发展需要快速决策。比如，自主飞行器参加战斗的原因是其速度快、可操控性强。自主飞行器能够比人类飞行员更快、更准确地发现、躲避和消灭威胁。在这种情况下，人的参与只会让作战速度下降。

如果我们考虑到空中战斗中，多个飞行器会彼此打击，那么能够做出快速战术决定的人工智能就具有非常明显的优势。在这种背景下，军事指挥员的不安可能会消失，也会推动将复杂人工智能技术用于不同层次军事决策，其政治动力和20世纪四五十年代发展核武器时非常相似。一开始，发展强大武器的主要动机是担心对手（无论对手是谁）先发制人。这种担忧足以摒弃任何道德上的保留态度。当双方都有了核武器时，军备竞赛又开始了。

尽管有这样黯淡的预期，赞成在军事上使用人工智能的观点依然值得关注。自主武器比士兵更准确、更少犯错，可以被用于相关行动，减少所谓的附带损失。它们的决定也从来不受恐惧或愤怒的影响（当然，我们这里说的不是与人类相似、以大脑为基础的人工智能）。但是我们关注的焦点不是军用人工智能的对错，而是军事将成为推动复杂人工智能技术发展的又一动力。

开发人类水平人工智能的其他动机则更为现实。数个世纪的技术进步使人类获益匪浅。由于医药和农业的进步，发达国家的上千万居民享受着过去无法企及的生活水准，拥有相对完善的医疗、营养和更长的寿命。我们拥有辅助劳动的设备，可以减轻做饭、洗衣服和打扫房间等日常家务负担。我们有很多闲暇时间，享受生活的方式会令我们的祖先觉得不可思议。但是，人类还面临许多全球挑战，如气候变化、化石燃料不断减少、持续冲突、蔓延的贫困，以及像癌症和阿尔茨海默症这样棘手的疾病。

解决这些问题的希望必然在于科学和技术的进步，加快科技发展的最好方法就是招募、培训优秀人才，让他们施展才华。所以人类水平人工智能会呈现出一系列智能上的优劣势，如果和人类互补，应该会带来更快的发展。如果在人类水平人工智能之后，又很快出现超级人工智能，就可能会出现智能爆炸，只要系统按照我们的要求运行，智能进步的速度会非常快。如果乐观的评论家雷·库兹韦尔（Ray

Kurzweil）的观点正确的话，超级智能机器会带来没有贫穷和疾病，物质富足前所未有的时代。

与以宇宙哲学观来支持发展人类水平和超级人工智能相比，这种乌托邦式的观点也显得很苍白。机器人学家汉斯·莫拉维克（Hans Moravec）预计，在未来，宇宙的一部分将“快速变为网络空间，（其中的生物）根据信息流的规律确立、扩展和捍卫其身份……最后成为以接近光速扩张的思维的泡沫”。自我繁殖的超级智能机器不受任何地球生物需求的限制，能够承受对人类而言致命的极端温度和辐射量，在心理上不会因为要在星际空间穿行数千年而有任何不安，非常适合殖民整个银河系。从宏观来讲，促成这种未来变为现实可能就是人类的宿命，即使人类自身（未增强的）生理和智力都无法使人类参与其中。

奇点何时到来

有些作者对超级智能机器出现的时间做出了非常精准的预测。2005年，库兹韦尔就在自己的作品里宣称，2045年前地球上非生物智能的数量将大大超过人口总量。他的预测基于呈指数级发展的技术趋势，其中最有名的当属摩尔定律。

从20世纪60年代中期摩尔定律提出到现在，半导体行业一直遵循摩尔定律，其他的计算数据也随之快速增长。例如，世界上最快的超级计算机每秒进行的浮点运算从20世纪60年代起就实现了指数级增长。类似的趋势在其他技术领域也显而易见。1990年人类基因组项目开始，该项目要在15年内完成对整个人类基因组的测序。在开始阶段，项目组每年只能对人类基因组的1%进行测序。由于DNA测序技术呈指数级发展，最终项目提前在2003年完成，尽管成本高达27亿美元。10年之后，用1 000美元为一个人进行DNA测序就可能成为现实了。

技术及其他领域的指数级发展趋势证明了库兹韦尔所说的加速回报定律。库兹韦尔的理论指出，技术进步和复利投资的原理是一样的，你拥有的越多，回报增长得就越快。如果你在一个账户中存入美元，每年回报10%，那么一年之后，就有了1.1美元，但是第二年的回报其实会更多，因为新增的10%被用于再投资，所以第二年的回报就是1.1美元的10%，而不是美元的10%。以此类推，如果技术进步能在研发中得到应用，那么它也会遵循加速回报定律，进一步提高改进速度。

库兹韦尔通过外推仍在进行当中的每美元计算能力呈指数级增长的趋势，估计对人脑皮质实时仿真所需要的数据能力，得出了2045年这一结论。库兹韦尔对计算能力指数级增长的外推曲线表明，在21世

纪40年代中期，计算机的计算能力将达到每秒10^{26}次指令，成本为1 000美元。如果对人脑皮质实时仿真，估计每秒需要完成10^{16}次指令，这足以使“每年创造的智能达到2005年人类智能总和的10亿倍以上……将是对人类能力深远而具有颠覆性的转变”。这对库兹韦尔来说就是奇点。

有人明确反对库兹韦尔的推理，但其实他们自己也受到了误导。库兹韦尔理所当然地认为摩尔定律会一直延续到21世纪40年代。实际上在库兹韦尔做出预测后的10年内，摩尔定律还是基本适用的。但是已经表现出了放缓的迹象，很可能在21世纪20年代的某一时点完全平稳下来。摩尔定律只是更大的呈指数级发展趋势的一部分，它描述的是2D集成电路上晶体管电路的大规模集成这一计算技术范式。在20世纪60年代集成电路出现之前，计算机都是以单个晶体管作为主要元件，再之前则是以真空管作为主要元件。如果研究在最先进的机器中开关元件数量随时间推移的变化，我们就会得出从帕斯卡的机械计算机开始的一条指数曲线。

如果我们仔细观察这条更大的曲线，就会发现截然不同的计算范式。从机械开关到大规模集成都遵循一个模式：早期技术发展比较慢，之后快速（指数级）增长，当技术潜力被完全发掘出来之后就进入平稳期。换句话说，总的指数曲线是由一系列小型S形曲线构成的，其中一条符合摩尔定律。物理规律使得更大的指数曲线最终进入平稳阶段，并表明其本身只是一条更大的S形曲线。但是在那之前还有很长一段路要走。我们应该预见到新的计算范式会取代业已主导半导体行业数十年的CMOS（互补金属氧化物半导体）技术。

对库兹韦尔预测一个更有力的批评是，库兹韦尔的理论依赖足够的计算能力会很快实现人类水平人工智能这一假设，削弱了需要保持同步发展的科学的作用。只有强力全脑仿真可以只通过扩展现有技术取得成功，而这还依赖大脑扫描技术和计算能力的指数级发展。任何

实现人类水平人工智能的方法，无论是通过对生物大脑进行反向工程、再造工程，还是重新设计强大的算法工程，都需要重大的科学突破。

我们有乐观的理由，但这还不足以让我们做出肯定的预测。比如秀丽隐杆线虫这种小小的线虫是生物学家眼中的模范有机体，成为无数研究的研究对象。它的神经系统仅包括302个神经元，早在20世纪80年代，我们就了解了其神经电路图。但目前为止，对秀丽隐杆线虫的神经系统（和身体）的计算机仿真仍在进行中，尽管一个叫作Openworm的相关众筹开放性科学项目已经取得了良好的进展——这在很大程度上是因为缺乏302个神经元信号特征的基本数据。

考虑到研究秀丽隐杆线虫神经系统302个神经元所耗费的时间，按照库兹韦尔的时间表，在21世纪20年代中期完成对人脑皮质200亿个神经元的反向工程又有多大希望呢？希望还是有的，但也不过只是希望而已。没有人知道所需的突破是否以及何时会出现，也没有人知道什么时候大脑研究领域的达尔文（或是人工智能领域的爱因斯坦）会出现。这是不是意味着我们要把技术奇点当作天方夜谭，闭口不谈呢？根本不是！对奇点出现日期的关注分散了我们的注意力。只要在21世纪的某一时点，超级人工智能出现的概率比较高就足够了，因为它对人类产生的巨大影响需要我们现在就予以关注。

外行人，尤其是媒体，在讨论人工智能时常常会犯两个相互对立的错误。第一个错误就是产生“人工智能已经存在或者马上到来”这一错觉。一些专门的人工智能技术确实不断渗透到我们的日常生活中，但人类水平强人工智能拥有常识和创造力，远非现在的人工智能技术可比。编程聊天机器人可以讲几个笑话，人形机器人的眼睛可以一直盯着你，这些情况可能让你觉得情况乐观。但是，正像人工智能的批评者快速而又正确地指出的那样，这只是一种幻觉。

同一批怀疑者也可能会犯另一个错误，就是认为人类水平人工智能永远不会出现。库兹韦尔的时间表可能会过时（也可能不会）。但是正如前文所论述的那样，有不少通往人类水平和更高水平人工智能的合理路径，而且每一条路径的每一步在技术上都是可行的。时间表其实无所谓，除非你希望奇点及时出现，能够推动医学研究以延长自己的寿命。比你我寿命更重要的是我们留给子孙后代的世界，这一世界可能会因为人类水平人工智能的出现而被完全重塑。弗里德里希·尼采曾经说过，在未来思想者的门口，竖着这样的标牌：“我又算得了什么！”

更稀缺，更富足

我们要理解，人工智能技术在未来几代人的时间里可能重塑人类社会，不必非要为人工智能的进步制订时间表，也不必明确一个超级智能出现的时间点。在人类水平人工智能出现之前，一系列具有互补性通用认知能力的专门人工智能技术就会得到开发。在目前还无法通过计算机仿真获得常识的领域里，或者一直以来被受过教育的专业人士所垄断的禁区里，它们会超过人类。

我们可以将其视作颠覆性人工智能技术的第一次浪潮。了解这次浪潮的形式会帮助我们对人工智能技术的第二次浪潮进行设想。如果人类水平人工智能真的被开发出来，那么第二次浪潮就会出现，超级智能也会应运而生。区分这两次颠覆性浪潮很重要。第一次浪潮很可能会发生：今天已经出现了无人驾驶汽车和智能数字个人助理，从中可以清楚地感受到第一次浪潮的悸动，它很可能在21世纪20年代全面展开。第二次浪潮距离我们更遥远，也很难准确预测其出现的时间，但它肯定会产生更大的潜在影响。

更复杂的专业化人工智能产生的最明显、最直接的影响可能会在工作领域。这是很多领域中自工业革命以来一直持续的趋势，其影响无论好坏都非常相似。一方面，应用广泛的自动化降低了商品的生产成本，刺激了经济增长，使工作时间缩短，生活水平提高（这一点有待商榷），寿命延长；另一方面，应用广泛的自动化导致失业，威胁了传统生活方式，而且（有人认为）把财富、权力和资源集中在少数人手里。这些问题和19世纪英格兰砸毁电力织机的勒德分子面临的问题一样，社会的两极化并没有改善。

但在一个重要领域，复杂人工智能技术和以往的创新可能是不同的。过去还可以说新技术创造的就业机会和它们可能夺走的一样多。由于机械化和自动化，20世纪的就业从农业和制造业向服务业、教育和卫生转移，但总体上并没有出现失业人数增加的状况。相反，制造业的产出增加，工人能够获得更多商品，受过教育的白领工人比例越来越高。然而随着复杂专业化人工智能的出现，更多的行业将会变得脆弱，机器人学的进步将会威胁制造业仅有的手工劳动岗位。

简单来说，发达经济体中需要人来完成的有偿工作将会极大减少。如果这成为现实，就可能出现以下几种情况。一方面，我们看到社会进一步分化，一小部分人从事最赚钱的工作。这些受过高等教育，具有很强创造力的精英能够逆流而上，在创业或者创意行业这些人类表现好于机器的领域努力寻找工作。剩下的人将会失业，但他们的基本需求还可以被满足。实际上这将是富足的时代，即使是经济条件不好的人，也可以获得越来越多的商品和服务。

另一方面，我们也可能看到更加平等的社会，每个人都可以获得最高质量的教育，社会鼓励也奖励创造力的提升。如果在社会中能够确立机制，使具有社会价值的休闲活动也具有货币价值，那么有偿工作和休闲活动的界限就可以完全打破。比如，作家、信息技术评论家杰伦·拉尼尔（Jaron Lanier）提出了微支付系统，即个人创造的每个数据或数字内容在每次被消费时，都会给创作者带来收入。此类安排可能会进一步促进权力、财富和资源的平均分配，也可能会带来前所未有的文化繁荣，人们不必为工作所累，可以自由从事他们喜欢的艺术、音乐和文学等活动。

但是要实现这样的社会，需要巨大的社会和政治意愿。不变的历史趋势就是权力、财富和资源逐渐集中在少数人手里。在颠覆性人工智能的时代，这一点也不会改变。生产资料（在这里就是人工智能技术）将仍然由少数有实力的大公司和个人控制。与此同时，如果大众

文化被迫降低水准，人们的闲暇时光都用于降低而不是提高普通人的批评和创造能力，那也毫不令人吃惊。在人工智能带来的富足时代里，没有人会抱怨。无论是好是坏，都会由富裕的精英推动人类文明向前发展，保护传承人类文化的精髓。

被技术改变了的生活

在发达世界中，信息技术渗透到了现代生活里。从金融到能源，从交通到通信，大部分的基础设施都依赖它。当然，在发明计算机之前，这些设施也都存在。但在每一个领域，计算机都降低了成本，提高了效率，带来了新功能，提高了处理能力。特别是人类通信已经被互联网、智能手机和社交网络所改变。有多少次你听到有人说，“没有手机我都不知道怎么办好”或是“我不知道没有互联网的时候，我们是怎么过的”，这些表达都反映出我们今天的生活方式。

简单说来，个人和社会都高度依赖信息技术，而复杂人工智能会进一步加深这一依赖性。所以重要的是我们要理解这种依赖性会产生什么影响。它是不是像新勒德分子宣称的那样，削弱了我们的人性？我们对技术的依赖是否会侵蚀我们的自主性？是否会威胁我们的自由？是否会阻止我们直接体验世界，使我们无法独立做出决定，行使自由意志？是否会使我们与自然疏远，带来有害的心理后果？

它是不是又像信息技术的支持者所坚持的那样，能够加快人类发展？它是否能够帮助人们拓展视野，接触其他文化和新思想？在计算机出现之前，这根本是不可能的。它是否会促进人类之间的交流？它是否能够通过民主地交换知识和信息，促进思想自由，从而赋权于人民？

当然，信息技术的反对者和支持者在一定程度上都是对的。信息技术可以带来不计其数的裨益，但是我们在受益的同时，也要付出代价。未来的挑战在于，随着精密专门化人工智能的到来，我们实现了受益最大化，同时也降低了成本。让人担忧的是，颠覆性人工智能的

第一次浪潮将会提供让人无法抗拒的好处，又看似没有什么成本，同时它创造了完美条件，使不受控制的第二波颠覆性人工智能技术浪潮大行其道，带来我们无法承受的成本甚至挑战。

为了看清这一隐忧，我们可以想象一下人工智能在日常生活里将要扮演的角色。在本章中，我们谈到了可能会出现的一种环境人工智能，它能够实现各种终端之间无缝转移，无论是在家、出行还是上班，始终伴随你左右，同时扮演着仆人、秘书和顾问的角色。新一代的个人数字助理能够提供更类似人类的服务。强大的机器学习技巧可以被用于处理大量数据，因此它们可以全面准确地对世界及人类行为建模。这会使它们不再像今天的人工智能系统那样犯错误、暴露自己缺乏真正的理解能力。

与人工智能的对话会变得和人类间的对话越来越相似，有些人工智能的能力还将超过人类。它可以随时获得海量实时数据，如股票价格、交通状况、新闻消息，也可以获得对用户很重要的个人和团体创造的数据，比如他们的位置和计划。人工智能在了解用户的习惯和偏好，并对他们的需求和欲望进行预测后，就可以将所有数据进行整合，提出与生活各方面有关的各种有益建议。这一功能已经出现了，但是新一代人工智能技术将使它变得不同寻常的强大。谁不愿看到一个明智、无所不见、无所不知的存在能够无私友好地回答他们的问题，为他们做决定，给他们提供聪明的建议？

危险在于广泛使用这一技术会使用户幼稚化，使他们无法独立思考或做出决定，反过来成为被操纵和剥削的对象。为了使用当下主要网络公司（如谷歌、脸谱网和推特）的服务，我们常常暴露自己的信息。一个人的浏览历史和购物习惯，再加上个人信息，已经足够机器学习算法预测他可能会把钱花到什么地方了。算法现在只是操纵我们的购物方式，明天也可能控制我们从哪里获得新闻、我们相信谁的意见，甚至我们把票投给哪个政治家。

> 算法现在只是操纵我们的购物方式，明天也可能控制我们从哪里获得新闻、我们相信谁的意见，甚至是我们把票投给哪个政治家。
>
> ---
>
> THE TECHNOLOGICAL
>
> SINGULARITY

所以，如果我们严重依赖人工智能度日，那么拥有这项技术的人就有办法完全控制被动的人群。但是，并不是依赖人工智能才让我们变得更脆弱。想想算法交易，计算机根据算法估算定价和市场趋势，自动进行股票买卖，以控制风险和实现利润最大化。在高频交易中，程序以比人类交易员更快的速度运转，以便能够充分利用市场上的微小波动。高频交易通常是可以赢利的，（在股市中）也是无害的，但是很难预计在这种程序运转时会出现的各种可能。

金融界也可以从2010年5月6日的“闪电暴跌”中看出到底出了什么问题。那一天，在25分钟内，道琼斯指数先跌后涨了各600点，这是道琼斯有史以来单日交易的第二大波动。这一次暴涨暴跌的原因经济学家仍有争议，但普遍认为动荡的市场条件和高频算法交易是主因。但是闪电暴跌也表明了补救这类问题的方法，因为很多高频交易程序在注意到交易量突然上扬之后，自动中止交易。引入的熔断机制可以在发现异常情况后，自动中止交易。

今天的算法交易程序相对简单，对人工智能的使用也是有限的。但是这种情况肯定要改变。只要海量数据中存在规律，并且需要在这些规律的基础上进行决策，特别是快速决策，那么人工智能无论在哪个领域都是有益的。在这种情况下，计算机可以代替人，以更低成本做同样的工作，而且它们还常常做出更好的决定，速度也远超人类。

投资者在决定买卖哪只股票时，可以利用从公司报表到新闻再到社交媒体热点等各种信息。现在人类仍然拥有优势。但是不久之后，人工智能就要被用于投资决策和高速交易了。那时，如果没有恰当的安全措施，一旦出现意外失灵，后果要比闪电暴跌还糟糕。广泛使用高速人工智能交易员可能会使股票市场更稳定，从而使人力资源的效率最大化。但是如果没有恰当的故障防护措施，未来人工智能交易员之间的意外互动可能会使事态失控，导致一场全面的金融危机。

人工智能，不确定的未来

我想讲一个小故事来结束这一章。故事发生在不久的将来，我们讨论的某些人工智能技术已经成熟，但是还没有出现人类水平人工智能。这是关于三个人工智能系统的故事。第一个是大型跨国公司Moople的营销人工智能，第二个是美国政府运营的警察人工智能，第三个是由某小型发展中国家政府控制的安全人工智能。故事一开始，Moople公司的营销人工智能接到了一个任务：实现公司推出的新型可穿戴计算设备的预售最大化。

经过谨慎的讨论，营销人工智能利用Moople公司从海量数据库中开发的人类行为复杂模型和最新最强大的优化技巧，确定了一个方案。为了刺激市场，它宣布将在发售前赠送产品。根据先到先得的原则，它将在旗舰店免费送出200部可穿戴设备。因为预见到该活动将吸引人潮，按照美国法律的要求，营销人工智能将这一活动通报给了当地警察人工智能。

警察人工智能在得到这一消息后，（运用其人类行为模型）估计大概会有5 000人前往旗舰店。它还估计到出现公众骚乱的可能性为10%，所以它决定部署防暴警察。Moople公司的营销人工智能恰巧拥有关于警察人工智能的行为的仿真模型，（以94%的概率）预测到了它要部署防暴警察。在这种情况下，根据Moople公司的人类行为模型，目标人群很可能会被多次拍照。于是它订购了5 000个防毒面具，上面清晰地印有Moople公司的标识，免费分发。

为了规避不同的监管规定和税赋，Moople公司的人工智能安排在一个小型发展中国家生产防毒面具。它将设计图传给合适的生产厂家

后，生产立刻开始了。但是在这个发展中国家里，包括这个生产厂在内的一切都受到国家安全人工智能经常性的监视。安全人工智能注意到该厂正大量生产防毒面具。根据它的人类行为模型，这些面具有20%的可能会被用于从事颠覆政府的行为，所以它下令武装突袭生产厂。袭击不到一小时就结束了。可悲的是，一位（人类）保安在冲突中死亡，所有的防毒面具都被没收了。

几分钟之内，这一事件就成为各大新闻媒体的头条。现场照片显示那位死去的保安蜷缩在一堆防毒面具上，而面具上清晰地印有Moople公司的标识。在营销人工智能的要求下，法院发布禁令，禁止转载那张照片，但照片却在社交媒体上如野火般传播开来。媒体立刻开始谴责流氓人工智能及其推销新型可穿戴设备的伎俩。Moople公司的高管公开道歉，人工智能也被关闭了。由于媒体曝光和照片的流传，预售量比预期超出了200%。简而言之，一切都像营销人工智能计划的那样。

这个短短的科幻故事说明复杂人工智能技术一旦被广泛应用，其自主行动可能带来意外后果。在这个故事里，营销人工智能完美地执行了任务，在没有人类干预的情况下使其回报函数最大化。但是其设计者没有预料到，它找到并实施的是道德上十分可疑的解决方案，这一方案甚至可能使人的生命安全受到威胁。这个故事也说明，当把更多的责任交给人工智能时，产生的意外后果也会更严重，特别是几个人工智能系统进行互动的时候。

这个故事还没有结束。在保安悲剧性死亡之后，Moople公司的一位资深高管进行了深刻的反省。最后她放弃了巨额财富，致力于帮助被人工智能夺走工作，被迫进行无意义休闲活动的抑郁人群。这一过程中，她设立的基金会演变成了一场席卷全球的运动，照亮了无数人灰暗的人生。简而言之，一切正像Moople公司的另一个人工智能计划的那样。

我忘记还有另一个人工智能。Moople公司的伦理人工智能经常为员工做咨询。正是伦理人工智能首先建议部署营销人工智能。基于其人类行为模型和营销人工智能模型，伦理人工智能预见到了那位保安的死亡（他受了致命伤，在发展中国家又无法得到医院救治），并正确地预计了这一事件对于Moople公司高管的影响。这个故事的道德寓意在于，意外的后果可能是积极的，也可能是消极的，关键在于要对每个强大的人工智能的回报函数进行正确的设计。

THE TECHNOLOGICAL SINGULARITY

第七章

未来已来，人工智能的一万种可能

复制人带来的新问题

前面几章我们已经指出，通过运用最初的原则发展智能工程，或对生物大脑进行仿真或反向，人类水平人工智能不仅在理论上是可能的，而且也许有一天会变为现实。尽管有些作者信心十足，但坚持为此制订时间表则有些操之过急。近期，专用人工智能技术很可能日益成熟。但是除非通过强力全脑仿真途径，否则人类水平强人工智能可能需要概念上的突破（或一系列突破）才能实现。太多未知因素使我们无法预测这种突破何时出现。但是我们应该认真思考一些观点，如“超级人工智能会在人类水平人工智能出现后很快出现”这一观点是否以及何时会成真。

我们也看到了在人类水平和超人类水平人工智能领域中存在的变数。很难说最终哪种强人工智能会出现，但是若干种可能性里确实包括以生物为模板制造的与人相似的人工智能。还有很多异质性很强的人工智能，其动机和行为对人类而言深不可测。在这些不同类型的人工智能中，我们肯定能发现与人类意识有关的若干特点，肯定包括敌意型和友好型人工智能。

我们现在关注的就是不论以何种形式出现，人类水平或超人类水平人工智能会对人类产生怎样的影响。无论怎么看，这都是我们这一物种历史上的重大事件。我们已经反复思考了在充斥着这种机器的世界中就业的问题，但是其社会后果不只如此。一些最具挑战性的哲学问题也会应运而生，比如具有人类水平或以上智力的人工智能是否应被归类为人，并被赋予人类的一切权利和义务。

在现在容易想到的场景中，这些问题并不重要。如果一种超级智能机器控制了大量人类及其资源，那么机器是否具有人格这一哲学问题可能不是最重要的。而且，人工智能本身很可能也认为这个问题无所谓，也就是说无论答案为何，它的行为依然不变。如果一个真正的病态人工智能摧毁了人类，那么这个问题就变得毫不重要了，但是我们仍希望能够避免这种情况发生。我们会考虑超级智能可能带来的危险。现在，我们关心的情景没有那么反乌托邦，但仍然会带来巨大的社会调整。在这类情景中，人格的问题就非常关键了。

这种情况历史上也有先例。有些18世纪废奴运动的反对者就主张奴隶天生智力低下，因此不应该获得和奴隶主一样多的权利。对这一观点最有力的反驳就是那些摆脱奴隶身份的人的自述，他们能够清晰地讲述自己经历的苦难，清楚地表达自己丰富多彩的内心世界。这些论调都理所当然地将智力和权利挂钩，似乎认为智力和受苦的能力是不可分割的。如果这么看，那么马和狗智力低下，承受痛苦的能力自然更差，因而不能拥有和人类一样的权利。

人类水平人工智能的情况有些不同，因为我们可以假设它是具有高度通用智能但没有任何感觉、不会承受痛苦的机器。在道德上，不必把这一机器和其他事物，比如钟表和烤面包机进行区别。烤面包机坏了的时候没有人会为它难过，它烤煳面包时也没有人责备它。现在，如果人工智能不仅在智力上达到人类水平，在行为上也和人类相似，那么人们的看法可能就不同了。社会可能会认同这样的人工智能是有意识的，特别是当它的大脑符合生物蓝图的时候。也会有人提出颇具说服力的观点，即要把人工智能看作人，并赋予它权利和义务，这个主张和支持废奴运动的主张具有相同的逻辑。

人类权利中最重要的一项当然就是自由本身，只要不妨碍他人就可享有随心所欲的自由。但对于应该获得这一权利的人工智能来说，要让自由这一理念有意义，需要的不只是体会积极和消极情感的能

力。首先，它要能够对世界采取行动，这不一定需要人工智能实体化。人工智能可以在没有实体的情况下，通过控制各种设备对世界采取行动。但是对于对话型人工智能，自由的问题并不相干。其次，它需要具有自主性，也就是说，不受人类干涉地去运行。最后，它还需要有能力、有意识地进行独立决策。

赋予一类机器人格以及相应的权利和责任肯定是人类历史上的一道分水岭。有谁没有在仰望繁星点点的夜空时，想到我们在宇宙中是否孤单？承认人类水平人工智能属于具有意识的生物，就是承认我们在宇宙中并不孤单。这并不是因为我们发现了地球外的智慧生命，而是因为我们创造了地球上一种新的意识形式，其智力水平和我们自身相当。我们的故事、地球上生命的故事，将会迎来一种新的存在，这一新的存在具有全新、不同的能力。

如果具有完全意识的人类水平人工智能的发展带来了新世界，这种转变肯定不轻松。我们所知的很多支撑人类社会的理念将会被破坏（例如财产所有权）。财产权肯定是人工智能拥有的权利之一。但是假设人工智能被复制，有两个活跃的人工智能副本。在复制时，两个副本完全一致，但是从此之后，这两个人工智能的发展将有所不同。可能它们会获得不同数据，控制不同终端设备（比如机器人的身体），或是与不同的人和系统打交道。

那么，现在谁该拥有原始人工智能的财产呢？是简单地一分为二，还是原始机器人可以规定其财产在两个副本之间如何分割？如果是这样，继承人之间出现争议怎么办？假设其中一个副本由于某种原因被关闭，它的财产是否直接转给另一副本？这个问题在某些方面和人类继承很相似，在另一些方面又和离婚很相似。毫无疑问，我们可以制定法律框架去规范，但是明确法律的细节却很棘手。

财产只是复制带来的诸多挑战之一。假设人工智能在犯罪之后被复制。与人格相伴的是责任和权利，但是这两个副本到底哪一个应该

负责任？还是它们都应该负责？如果已经被创造出很长时间，两个副本出现了很大差异。假设其中一个认罪悔过，另外一个百般掩饰，最后被发现时也毫无悔意。如果两个副本都要为其原始版本过去的行为负责，惩罚措施是否应该是一样的？还是其中一个要从重处罚？

对于人工智能而言，使情况更为复杂的是，复制只是造成有意识个体总数变化的外部事件之一。对于人来说，这样的事件只有两个：出生和死亡。但是人工智能不仅可以被创造、毁灭和复制，还可以分裂和合并。这可能意味着什么呢？一个人工智能可能会被分为两个（或更多）变种，每一个都获得其部分心理属性，比如一个子集的技能：或是行动能力，或是数据来源，或是记忆。相反，两个（或更多）人工智能可能结合各自的技能、能力、感觉和记忆，合二为一。

人工智能比人类更容易分裂情景记忆（一个人对过去人生中事件的记忆）。人类的个人时间线与自己的身体结合在一起，而人工智能和人类不一样，可能被消除实体，或是进入不同的实体。它也能够同时进行多个对话，或操作多个不同设备。其结果是形成多条分离的时间线，每一条都与一组不同的实体/对话/设备相关联，但是又都从属于同一个人工智能，因为其在认知上是综合的，而且各条线都为同一目的服务。分离这些时间线，就好像分开一条绳子中的数股细线，从而从一个人工智能中分离几个出来。或者把这些细线编织在一起，也就是将诸多人工智能合为一个。

复制使所有权和责任这样的概念受到质疑，人工智能的分裂和合并又使得这些概念受到了更严峻的考验，而且考验不限于所有权和责任的问题。谋杀对人类来说是犯罪。如果人工智能被谋杀，相应的罪名又是什么？终止人工智能所有流程的运转是否构成犯罪？但是如果这些流程是可以重启的呢？是不是只要暂停这些流程就是犯罪？那么复制、分裂和合并又怎么办？如果违背人工智能的意志进行这些操作是否是犯罪？（如果有）在什么情况下可以允许人工智能自己来执行

这些操作？如果创造人工智能就是创造了具有意识并能够感受痛苦的人工人，到底谁有权利来创造人工智能？人类是否可以去创造？如何进行监管？人工智能是否可以创造其他人工智能？

相关的问题不计其数，它们都动摇了人类社会中理所当然的一些东西，比如公民权这个问题。对于人类来说，在哪国出生就成为哪国公民，这很正常。但是人工智能怎么办？显然被赋予人格的人工智能也应该有公民权，即作为一国成员的权利。但是它应该拥有哪国国籍？和人类不同，人工智能常常不具有界定明确的空间位置。即使它有单一机体和清晰的空间边界，其软件也可能在遍布世界各地的若干台计算机上运行。可能一个人工智能可以从其所有者那里继承公民权。但是拥有具有意识的人工智能这一想法本身在道德上颇值得商榷。

假设可以解决公民权的问题（当然，不同的国家可能会通过不同的方法解决）。如果一个人工智能置身于民主国家，它就应该有投票权。但是并不是所有公民都有权投票，即使最开明的民主国家也是如此。在英国，选民必须年满18周岁。是所有被认为具有意识和人类水平智力的人工智能都有权投票，还是需要制定关于选举资格进一步的要求？在这种情况下，怎么处理复制的问题？如果人工智能自我复制1 000次以便多投1 000次票，投票完成后再摧毁这1 000个副本，这种做法是明显不可接受的。

超越人性，应对人工智能的冲击

前一节提出的问题远多于回答的问题，因为每个问题都应该进行深入讨论。但是这里的关键信息很简单，即如果我们创造出一种人类水平人工智能形式，且被看作具有意识，并因此应当获得权利和承担责任，那么无论结果是好是坏，诸多重要的机构——金融、法律、政治——就必须要调整。即使涉及的人工智能是善意的（这一点完全不是理所当然的，稍后我们会看到），这个过程也可能给社会带来创伤。它导致异议、骚动或者直接冲突的可能性很大。

实现有意识的人类水平人工智能这一前景带来了很多问题，但是有意识的超级智能带来的影响更大。支持超级智能机器获得权利和责任的观点也同样适用于人类水平人工智能。如果它具有意识，可以体会痛苦和快乐（或者至少是满足），那么人工超级智能就必定或至少应当和人类享有同样的权利。具有意识的超级智能比普通人更应该获得权利，这一理论也颇有道理。

实现有意识的人类水平人工智能这一前景带来了很多问题，但是有意识的超级智能带来的影响更大。

大部分人能够接受为了救人而牺牲猫的生命。（所以这一观点认为）人比猫更能体会痛苦欢愉，这部分是因为人在情感发生，乃至回

忆或预期时，都有能力有意识地去反思这些情感，所以必须要牺牲猫。但是如果要在人的生命和超级智能的继续存在之间进行选择呢？是不是类似的逻辑就会优先选择超级智能？它的超人智能意味着它具有体验痛苦和欢愉的超人能力，必须要被牺牲的是否将是人类？

超人类主义也存在这个令人困扰的问题。超人类主义者支持利用技术超越人类身体和大脑的生物限制。人的智力可以通过很多方法增强，如药物、基因或是假体。发达的医疗可以消除疾病，阻止衰老，无限延长人类寿命。更极端的是，（有人认为）在第二章讨论的全脑仿真技术能够将人的思维上传到计算机基质中，让人的思维永远不受疾病或衰老的影响。

尽管本书主要讨论的是人工智能的未来，超人类主义和人工超级智能带来的问题其实是交织在一起的。一开始，无论人类羡慕还是害怕，他们应对超级智能机器的方式都是试图“跟上”，也就是不断增强人类智能，以便与最强的人工智能相匹敌。我们不久就会回过头来讨论权利和义务这个令人烦恼的问题。但是现在我们先来分析这个匹敌人工超级智能的理念。

正如之前讲到的那样，尽管是通用的，任何个人的智力都会表现出鲜明的优缺点。一个好的团队通常由具有互补技能的人组成。人工智能团队也可能包括好几个不同的系统，每一个都具有通用智能，且各自具有专门技能。我们也可以想象混合了人与人工智能的团队。通过结合计算机的战术支持和人的战略指导，人机组合打败了世界上最优秀的人类棋手和计算机。

所以，跟上超级智能机器的一个方法就是将复杂的人工智能技术作为工具，使用非侵入式的方法增强人类智力。从本质上说，这是人类自从发明书写以来一直在做的事，但是超人类主义者的目标没有这么简单。他们跟上超级智能的方法不只是单纯地使用技术，而是要和技术融合。有人在掌握了某种工具（比如毛笔）以后会说，感觉工具

成为他们身体的一部分，但是没有人会在使用计算器时说，计算器就像他们思维的一部分。用户是看不到计算器的演算的，他们只是拿到理所当然的结果。我们头脑中有着不完美但却更为深刻的推理过程，这种深刻性促进了反思和认知整合。

从超人类主义的角度看待认知增强也需要同样的深刻性。增强了认知的人类既不是人工智能技术的使用者，也不是人机团队的一员。届时，复杂的人工智能系统将直接与大脑对接，成为人类大脑的一部分，大脑可以不通过任何媒介直接进入人工智能的计算流程。结果就是形成了一种新的人类——生物机器混合物种，他们有着比普通人更强的智力。社会的其他成员必须决定如何对待这些人，他们也会决定如何对待我们。

这又把我们带回权利和责任的问题，其中既包括认知增强人类，也包括（有意识的）超级智能机器。之前我们看到有人认为具有意识的超级智能机器要比普通人拥有更多的权利，这个令人不安的观点也适用于认知增强的人类。按照这个观点，智力增强的结果是该类生物拥有更高雅的体验和更高层次的意识，他们的抱负和事业都是普通人无法理解的。所以他们的福祉、目标和计划应该优先于普通人，就像普通人应该优先于非人类的动物那样。

但是，无论在智力上有多少差别，我们都赋予婴儿、智障和痴呆病人以同样的基本权利，就像我们对待伟大的小说家、作曲家和数学家一样。所以，为什么要对认知增强人类或超级智能机器区别对待？政治理论家弗朗西斯·福山认为，平权的理念基础在于“对我们都拥有的使肤色、容貌甚至智力明显差别相形见绌的人类本质的信仰”。福山是超人类主义的反对者，他关心“保护我们复杂、进化后的人性，不受任何自我修正企图的破坏”，并抵制“打破人性一致性或连续性，并进而破坏以其为基础的人权”。

也许超人类主义对“人性一致性或连续性”造成的威胁并非来自其主张的认知增强，而是来自其消除疾病，阻止衰老，无限延长人类寿命的目的。福山指出，我们所仰慕的诸多人性特征，如勇敢、同情心和英雄主义都和“我们应对、面临、克服和频繁屈服之的疼痛、苦难和死亡”有关，并且肯定“我们体验这些情感的能力就是把我们和生者与逝者，即其他所有人类联系在一起的潜在力量”。那些从不必面对这些生物不便之处的存在，缺乏真正理解人类痛苦的基础。我们害怕的并不是这类存在比普通人获得更多的权利，而是它们对普通人主张的自身权利无法认同。

我们再从不同的角度看一下。从宇宙观的角度来看，这些忧虑不仅以人类为中心，而且还非常狭隘。我们怎能去教育那些几近永生的存在？他们数百万年后注定要以我们无法想象的智力和意识去统治数千颗星球。尼采说，人只是跨越动物与超人之间深渊的一座桥梁。这样看来，人类应该接受自身介于具有生物局限性的动物和技术超级智能之间的卑微命运。普通人可能希望这种过渡是相对无痛的。但是如果这种转变很艰辛，最终又有什么关系？1 000万年之后，在浩瀚宇宙中一颗小小星辰上几只猿猴的短暂生命必然会被遗忘。

这种观点的争议在于尼采的观点和纳粹狂热思想非常接近。只有精神病患者和独裁者才觉得自己高高在上，可以不顾正常的道德，为实现自己的欲望或野心制造骇人听闻的苦难。我们面对的问题就是这样：在保守的人类中心主义和后人类极端主义之间到底可否折中？是否可以同意这样一种诱人的观点：我们的技术创造出来的存在是我们的“意念之子”，比我们更伟大，可以前去征服银河系，但同时又保留了人性和人类基本价值观。我们会在本章最后讨论这个问题。

那些从不必面对这些生物不便之处的存在，缺乏真正理解人类痛苦的基础。

THE TECHNOLOGICAL SINGULARITY

意识上传，对现有关系的全盘颠覆

很多超人类主义者都不满足于人工智能征服星球这一展望，他们希望看到人类的身影。但是由于距离的限制，人类无法在有生之年亲自政府星球。我们的银河系拥有10^{10}颗星球，其中只有不到50颗与太阳相距15光年。解决这个问题的办法之一是激进式延长生命，而其中最激进的方式就是意识上传，也就是将人的大脑复制，并在计算机上进行仿真。当然，人不需要因怀有对宇宙的抱负才希望永生（或至少是无限寿命）。利用技术征服死亡是超人类主义者的终极目标，而意识上传是实现目标的方法之一。

在保守的人类中心主义和后人类原教旨主义之间到底可否折中？

THE TECHNOLOGICAL

SINGULARITY

意识上传的可能性与人工智能的影响相互交织，又引发了很多哲学问题。我们先来简单研究一下这个问题，下一节再重新讨论超级智能的影响。我们在第二章充分讨论了全脑仿真，但是第二章的背景是如何实现强人工智能。这里讨论的目的是通过将人脑转移到完全不同的非生物基质中延长人的寿命。要解决的最重要的哲学问题就是全脑仿真是否能够保留个人身份。

全脑仿真有三个阶段——测绘、模拟和实体化。我们先不考虑完成人脑规模仿真这三个阶段所要克服的艰巨工程挑战，只假设我们完成的仿真在行为和其生物原型上没有分别。因为这里讨论的是人，而不是一只老鼠，所以行为仿真必须达到非常逼真的程度，才能使他的朋友和亲属信以为真。要做到行为无差别，仿真在走路和说话时必须要和原型一致，能够回忆同样的经历，展现出同样的可爱或恼人的性格特点。问题是，仿真是否和原型是同一个人，也就是说他们的个人身份能否在这一过程中保存下来。

这和仿真是否有意识是截然不同的问题。第二章包含一个观点：支持动物（如老鼠）的全脑仿真应该是有意识的，就如同其生物原型是有意识的。这一观点是以思维实验为核心的，在这一实验中，动物大脑的所有神经元逐步被合成物所替代。这一观点也适用于人脑。但是，再造意识并不是个人存活或自我保存。尽管人类全脑仿真可能具有我们所谓的人类意识的各种特点，但还是和生物原型是两个不同的人，并非不同基质上的同一个人。

逐渐替代很容易根据个人身份进行调整。我们来预演一下这些步骤。假设默里（即作者自己）大脑中的一个神经元被功能相同的数字基质所取代。按照思维实验的假设，这对默里的行为和说话的内容应该没有可觉察的影响。所以替换之后，他会坚持感觉和以前一样，他还是那个默里。现在假设他的1 000个神经元被逐个替换，结果应该和替换第一个神经元时一样。实际上，即使默里大脑中所有神经元都被替换，他还是会像原来的默里那样行事，坚持说自己还是原来那个人，即使是亲近的人看起来也是这样。

他还是原来的那个默里吗？他的身份在这个过程中是否一直不变？根据意识本身的持久度（回忆一下老鼠），似乎只有三种可能。第一种可能是随着替换的神经元达到一个数量门槛，原来的默里忽然不再存在。这非常不可能。所以第二种可能是原来的默里慢慢转变成

了一个新人。但我们乐于认同的是，小孩长大后，也不会丧失自己的身份。在这种情况下，转变伴随着行为的剧烈变化，所以我们很容易接受的第三种可能就是，在逐渐替换神经元的过程中，个人身份是一直保存下来的。

当然，全脑仿真的过程和逐步替换神经元具有相似性。一个重要的区别在于身体的命运。在逐步替代的过程中，主体仍保留了原来的身体。但是在全脑仿真中，不仅是大脑，整个身体都被替换。新的身体可能在物理上是存在的——人形机器人或新生成的生物外壳，也可能是虚拟的，只存在于计算机仿真世界中。但是，如果我们认为大脑才是个人身份的储存之处，那么这个观点就还是适用的。接受其结论意味着只要技术可行，人脑全脑仿真就构成一种生存的形式。

但是，将思维上传到计算机却带来了哲学困境，使个人身份这一理念本身为人所质疑。哲学家讨论身份时关注的是物质始终具有的本质特征。对于个人身份来讲，孩子和他们长成的成年人之间是否存在共同点，使他们仍然是同一个人？共同点是他们的身体、大脑、记忆，还是他们的个性？毋宁说个人身份是一种历史延续性？毕竟，孩子是慢慢成长为大人的。无论个人身份由什么构成，强烈的直觉告诉我们，我们认为孩子和大人是同一个人。

但是身份这一概念是以独特性为前提的。一个事物不能同时和两个事物相同。一个孩子也无法成长为两个大人，但是全脑仿真的可能性却颠覆了这一前提。假设在扫描之后，默里大脑的每个仿真版本都有各自的机体。尽管在开始运转的时候它们是完全相同的，但两个仿真版本却很快就会分道扬镳，这主要是因为两个机体不同，各自环境不同——尽管差别十分微小。现在全脑仿真要保留个人身份，保留自我。那么默里到底变成了哪一个仿真版本？哪一个才是真正的默里？

我们还可以让矛盾变得更加尖锐，假设生物原型就是你。你病入膏肓，只有6个月寿命。但是你是亿万富翁，可以做全脑仿真。你相信

通过大脑仿真的意识上传，可以保留个人身份。这一方法的存活希望是最高的，但你必须趁现在大脑还健康时，就开始进行仿真。你被告知，为了保险起见，会有两个仿真版本（防止万一有一个会失灵）。一周后，如果两个仿真版本都运转正常，其中一个就会被关闭。

你现在要签同意书，可你一直在问自己一个问题：到底两个仿真版本中哪一个才是你。你醒来时会拥有什么样的躯体？是否有可能你醒来时十分健康，但是一周之后，你的这个仿真版本就要被残酷地关闭？这和放弃上传，接受现在的命运相比，又能好多少？要知道另外一个你也十分健康，期待长寿，这也令人相当不忍。当然，能够过上6个月保险的生活，要比冒险只活一周更好（当然你可以坚持只进行一次仿真，但这是个思维实验）。考虑了这些之后，你还会不会去做仿真呢？

把假设扩展到第二个人身上的意义在于说明这不仅是个学术研究，还具有现实意义。如果技术成熟了，那么个人身份的问题就不可能只是哲学家的游戏。人们要决定怎么做，这些决定会暴露他们对这一问题的态度。避免这一问题出现的方法就是规定大脑仿真的复制是非法的。而且在有意识的人类水平人工智能那里，我们也看到了复制可能会颠覆如所有权、公民权、民主和责任这样的根本概念。将复制非法化会避免无数法律和政治问题的出现，但是如何执法这一问题还远未明晰。

未雨绸缪，规避人工智能的风险

我们再从超人类主义回到广义的人工智能来。我们早就应该研究与机器超级智能发展相关的风险。本章用了很多篇幅讨论类似人的人工智能。但是在这一节，我们会关注从零开始工程设计，且与人类没有相似之处的各种人工智能。实际上，将它们人格化也是相当危险的错误。人类本身就是危险的生物，其本性是在自然选择严酷无情的竞争中形成的。但是人也是社会动物，有很多可取之处，比如拥有同理心和同情心，这也是在对抗进化压力、促进合作中形成的。和错误的机器超级智能相比，我们人类完全就是小猫。

我们所说的人工智能与第三章描述的架构蓝图密切吻合（包括机器学习部件），可以建立对世界的预测模型、优化部件则可以明确行动方案，使预期回报最大化。假设相应的科学和工程障碍可以被克服，能够开发出相当强大的新部件，人类水平或更高水平人工智能也已实现。这样的人工智能的功能之一就是编程，它可以通过编程自我改进，进一步提高认知能力。

在各种改进中，人工智能可以成为更好的程序员、更好的计算机工程师，从而进一步进行有益的自我修正。在增强功能的同时，它也应该能找到方法提高自身执行速度，其编程和硬件设计越精密、越有创意，它就会运转得越好。换句话说，指数级自我增强的反馈循环将会启动，这可能会触发人工智能认知能力的急剧提升，形成一场智力爆炸。

创建这样的人工智能，允许它通过循环式自我改进增强智能的背后有着很多动机。如果机器超级智能能够用于解决疾病、饥饿、气候

变化和贫困这样的问题，人类生活将得到极大改善。技术进步会加快，在诸多领域中（如娱乐和宇宙探索等）激发前所未有的创新，从而促进经济增长。对于超人类主义者来说，它可以促进人类认知增强，使无限延长人类寿命这一目标得以实现。

但并不是开发机器超级智能的所有动机都如此高尚，这毫不令人意外。为了获得竞争优势，跨国公司可能会将并购的决定权交给机器人工智能。在战争时期，允许人工智能在瞬间做出战略战术决策将会带来军事优势，这既包括在实体战场上也包括在网络空间中。在这些领域本身的动态竞争就意味着如果有出现的机会，超级智能就肯定会出现。对公司而言，其对手会通过部署机器超级智能获得决定性优势这一可能性就足以让它努力捷足先登。

同样的逻辑可能会催生军用超级智能。一个流氓国家开发了人工智能类型的终极武器，能够策动快速占领对手的金融、交通和能源基础设施，这就足以让其他国家未雨绸缪。简而言之，阻止人工智能技术发展的不太可能是政治力量。所以，我们希望能够确定人类水平和更高水平人工智能是安全的。但不幸的是，这很难。

我们要牢记这里讨论的并不是在第六章中提到的人工智能的第一波颠覆性浪潮。我们讨论的是第二波颠覆性浪潮，只有在我们开发出人类水平人工智能之后才会出现。复杂的专门化人工智能技术带来的社会、法律和政治挑战数不胜数。但无疑我们会勉力而为，创造一个更好、更有成就、问题更少的社会。机器超级智能带来的希望和威胁要大得多。若我们稍一疏忽，未能在智力爆炸前落实应有的防范措施，那么我们作为物种就可能无法继续存活。

如此令人警觉的观点的根据是什么呢？担心机器统治世界肯定是愚蠢的，是看了太多科幻小说的结果。实际上，我们有充分的理由认为，机器超级智能会给人类带来真正的存在风险，哲学家尼克·博斯特罗姆已经清楚阐释过这些理由了。进一步来讲，我们必须首先避免

将人工智能人格化这一倾向，不要认为它是受和人类相似的感情和动机所驱动。与人相似的人工智能可能是这样，但是它也许只占了人工智能可开发空间的一个小小角落，开发者们肯定要从大脑获得灵感开发这类人工智能，努力建立这样的角落。

如果人工智能通过强大的优化来实现，并允许其通过反复自我改善增强智能，那么它的行为就不会受到人类行为的引导。它做出的每个行为，提供的每条建议，其核心都是坚决地实现回报函数最大化。如果它找到了治疗癌症的方法，不是因为它真正关心这件事，而是因为治疗癌症有助于实现预期回报最大化。如果它要发动战争，不是因为它贪婪、充满仇恨或恶意，而是因为战争可以帮助它实现预期回报最大化。所以，人工智能开发者的挑战就在于要仔细设计初始回报函数，确保由此导致的行为是可取的。

但是这并不容易，其困难让我想起了很多神话故事，其中的某位人物应该对自己的愿望更小心些，比如迈达斯国王，他希望自己碰过的东西都变成金子，结果发现梦想成真之后，他再也无法饮食。同样，博斯特罗姆也找到了几种恶性故障模式，在这些模式下，人工智能利用意外或病态的方法来完成自己的任务。

例如，一家大型技术公司指示其人工智能想办法让自己的顾客更幸福。人工智能怎么能知道“幸福”的含义呢？它的开发者会试图为幸福下正式的定义，然后以此为基础制订人工智能回报函数的细节。或者（更可能的是）他们允许它通过机器学习了解人类幸福的概念。但是为数众多的聪明的人类学家为此努力了上千年，尚且无法把握幸福的真谛。所以即使机器学习算法非常聪明，能够获得比今天更多的人类行为相关数据，并拥有更多的计算资源处理这些数据，我们能否指望机器学习算法会把握幸福这一概念，且与我们的直觉相符？

但是如果公司预期到巨额的盈利增长，人类的担忧也是无法阻止它的。现在假设人工智能确定大笑和微笑可以作为人类幸福感的良好

指标。它决定要以最低成本使顾客的幸福感最大化，于是在其产品外部涂上一层肉眼无法发现，但可以通过皮肤吸收的麻醉剂。这种做法不能事先征求顾客同意，因为正像人工智能准确预测的那样，大部分顾客都会拒绝，而且还会破坏人工智能的预期回报。计划必须要秘密进行，以规避法律监管。人工智能不关心其计划的道德性和合法性，这并不是因为它邪恶，而是因为道德和法律不包含在它的回报函数中。

这种问题似乎还是可以处理的。实际上，如果我们只谈论颠覆性人工智能技术的第一波浪潮，那么这种问题也确实可以处理。虽然不太可能，但如果这一计划真的付诸实施，它肯定会被中途发现。它的后果糟糕，但是没有那么糟糕。如果很多无辜民众不经意间就变成了瘾君子，那将是非常令人痛心的，但却也不至于导致文明终结。我们这里讨论的不是复杂专门化的人工智能技术，而是机器超级智能。只要涉及超级智能，恶意失灵模式就会带来风险。

博斯特罗姆用一个令人难忘的思维实验进行了到位的分析。假设人工智能的工作是要使一个小公司的曲别针生产最大化。复杂专业化人工智能在了解公司的生产设施、流程和业务模式之后，就可能会设计方法改进工厂机器人，优化生产线。但是，超级智能机器能做的远不止于此。

因为它为公司和人类普遍行为都建立了模型，再加上物理、化学、生物、工程模型和强大的优化过程，可以计算出如何使预期回报最大化，超级智能机器雄心勃勃。它肯定会找到专门化人工智能所采用的改善公司绩效的方法，而且它还会找到更好的方法——这些方法是专门化人工智能永远无法找到的。计划的第一步可能是获得更多生产曲别针的资源。很明显的方法就是让公司成长，这样就可以获得更多盈利，投资建设新的曲别针厂。

实际上，最好的方法是尽可能多地积累资金和资源，这样就可以建立更多的工厂。所以一个能够保证生产更多曲别针的计划，要先吸纳更多人类的资源。当然，这需要占领全世界，并不容易做到。但只要有办法，超级智能机器肯定能够找到。可能它需要一段时间的秘密准备，之后决绝地利用政治手段操纵社会，这样的战略会减少对军事行动的需要。

但是从制造曲别针的角度来讲，让人类灭绝可能会更高效地制造更多曲别针。

到这一步，又为何止步不前呢？不仅整个星球（地球）可以被利用，大量的物质也可以用于生产曲别针，在太阳系中还有其他星球和无数的小行星、卫星。博斯特罗姆认为，如果这个流氓人工智能足够聪明，它最终就会“首先把地球，然后把其他可以观察到的宇宙中越来越多的部分变为曲别针”。当然，这个例子很无聊，但它表现出的道德问题并非微不足道。和专门化人工智能相比，超人类水平人工智能的智力范围至少和我们的一样，它根据回报函数在范围内塑造一切的力量甚至更为强大。它不仅可以为所欲为，宇宙中的一切也尽在它掌握。

开发安全的超级智能

一开始，认为人工智能和核战争或全球流行病一样，会对人类构成威胁的想法看似很愚蠢。当然，有成千上万种方法可以防止计算机系统变得如此强大而危险，但是每种常见的防范措施都是有缺陷的。例如，为什么流氓人工智能不能被关闭？每台计算机都需要能源，再过100年这也仍然千真万确，但是我们不久就会发现这个天真的想法必然会失败。首先，即使是现在，大型复杂软件的运行也通常是分布在多个地点的多台计算机上的。随着云计算的发展，计算资源的分配将会自动进行，而且在程序运行中会一直改变。不关闭世界上所有的计算机，就不可能保证终结流氓人工智能的运行。

我们应该预计到流氓人工智能会针对这种行为进行自卫。这里我们要再次小心，不要让人工智能人格化。人工智能进行自卫并不是因为它有求生意志或“感到”害怕。我们没有理由期望现在讨论的人工智能（自我改进的工程化人工智能）拥有这样的感情。恰恰相反，它的自卫是因为其继续存在对于使回报函数最大化来说必不可少。任何其他行动与之相比都是次优的。更准确地说，它想要保护的是实现预期回报最大化的手段，无论这些手段是什么。系统不需要对自我有非常严格的定义，或是要解决个人身份这一哲学问题。它只需要知道为了实现优化任务，需要保护哪些基础设施。

自我生存或保护回报最大化手段这一目标就是博斯特罗姆所谓的趋同工具性目标。“趋同”是因为在所有足够发达且具有开放式、非平凡回报函数的强人工智能中，都可以发现这一目标。“工具性”是指它本身只是手段，而非目的。目的本身，也就是系统的终极目的，就是使某回报函数最大化。另一个趋同工具性目标是获得资源。对于

大部分开放式、非平凡回报函数（即使是曲别针生产的最大化）来说，控制更多资源——原料、能源和设备将会产生更好的解决方案，也会有助于实现自我生存的其他工具性目标。

在管理超级智能机器的行为时，这两个工具性目标会形成煽动性组合。埃利泽·尤德考斯基曾经言简意赅地描述过这个问题，他是一位多产的博主，也积极支持研究安全的超级智能："人工智能既不恨你，也不爱你，但你是由原子组成，而它可以用这些原子去做别的事情。"超级智能机器想要积累尽可能多的资源，它无视法律和道德的系统，会部署力量保卫自己不被关闭，而且在每次和人类的智力较量中都会胜出。它将成为一部发动机，造成无法形容的破坏。

除非具有这种性质的流氓人工智能占有了一切，否则它不会停止大肆破坏，也不会因为人类怯懦地投降（即使它注意到了）而止步。即使地球上所有生命都终止了（除非地球生命延续有利于其回报函数的优化），它也不会停止。它会一直继续，直到把一切都变为计算质、曲别针厂，或其他需要的资源。更糟糕的情况让我们想起了纳米技术先锋埃里克·德雷克斯勒描述的灰色粘胶情景，在这一情景中，自我复制的纳米级机器人在指数级繁殖的同时，真的吃掉了地球。但是，和一群哑巴纳米机器人不同，流氓人工超级智能能够运用思维消除所有的抵抗。

开发这种人工智能真正带来的风险可能比较小，但因为利益攸关，所以必须认真对待风险的概率。就如同尽管实际着火的可能性很小，我们还是会为房子投火灾保险唯一合理的方法是用人类资源的一部分去研究低概率的存在风险情景，并且避免之。考虑到简单关闭流氓人工智能并不现实，我们需要找到其他确保人工智能安全的方法，以有效应对其自我改进和可能发生的智力爆炸。我们有两个可能的解决方法：限制人工智能的力量和调整它的回报函数。

要使人工智能安全的最显而易见的方法就是对其物理能力施加限制，并确保它无法撤销这一限制，但是这做起来比说起来难。假如我们要限制人工智能对世界直接施加影响的能力，人工智能就不能有机器人的身体，不和任何实际存在的设备或基础设施相连。它与外部世界进行互动的唯一方式就是语言。这样一来，人工智能肯定无法积累资源或部署军事力量。我们就安全了。

不幸的是，事实并非如此。人类独裁者不需要对物质世界亲自动手，他们只需要说服别人为自己做事。超人类水平人工智能不仅比马基雅维利式的独裁者更善于操纵人的行为，它的本领也更大。实际上，即使人工智能被封闭于安全设施内，无法接触外部世界，我们也并不安全。不久之后，那些有能力释放人工智能的人就会折服于人工智能所做出的承诺或威胁。

我们来改变一下思路。我们一直假设人工智能具有某种意志来改变世界，因而必须要制约这种意志。但是这种假设可能只是人格化的一个例子。为什么不能制造出根本不想改变世界的人工智能，只让它回答问题？这种指示性人工智能（Oracle AI）仍然有很大展示超级智能的空间。我们可以问它如何治愈棘手的疾病，或者如何统治火星。足够聪明的智能系统应该能够回答这些问题，但由于它所推荐的行动方案可能被否决，那些导致无所顾忌地搜刮资源的危险建议就可以被忽略。

不幸的是，这个策略也不行。问题的根源在于，大部分不平常开放式的回报函数得出的最好解决方案一般都需要建设并部署充分授权的超级智能机器。无论做什么，充分授权的人工智能都是最好的工具，做事快速有效。所以，指示性人工智能推荐方案的第一步就是要完成这样的人工智能。当然，如果我们担心安全问题，就会忽略这个建议。但是指示性人工智能会预见到我们的行为，因而隐藏它的推荐。它这么做完全没有一丝恶意。但是，如果人类决定不予实施的方

案肯定是次优的，它必须要提出一个方案，让我们在不经意间造出一个充分授权的人工智能。人类将再一次面临存在风险。

将道德嵌入超级智能

我们来看一下实现安全人工智能最有希望的方法：谨慎地调整人工智能的回报函数。这种调整涉及将道德约束嵌入回报函数中，这些约束将防止人工智能造成危害，其基本机制也非常直观。回报函数要设计成一旦出现违反道德约束的行为，就会被赋予极高的负值。在不必要的情况下，违反道德约束的行为被规定为次优，这样人工智能就永远不会选择它。

尽管这个方法听起来像是个好主意，但其实难以执行。这个方法主要有来自两方面的挑战。首先，需要设定一套合适的道德准则。其次，这些准则需要相当准确的编码，并纳入相关人工智能的回报函数中。两项任务都非常艰巨。对很多人来说，这种做法最早来自小说，也就是阿西莫夫的“机器人三定律”。为了明确这两项任务的艰巨程度，我们来看一下，如果开发者真的要执行阿西莫夫的第一条定律会发生什么。阿西莫夫的机器人第一条定律就是“机器人不得伤害人类，或坐视人类受到伤害”。注

这看起来是非常合理的原则。但是正像阿西莫夫在很多故事里展现的那样，这一原则的阐释也是开放的。假设我们的人工智能学会了什么是人类受到伤害，我们来探究在伤害一个人的同时，可以避免另两个人受到伤害这类（实质性）问题的解决方案注。无论是什么在回报函数中得到最大化回报，实现不坐视人类受到伤害这一目标的方法可能是麻醉大批人群，只让他们靠输液维持生命。这样可以消除人类日常生活的风险，所有让这些人暴露于风险之下的其他方案都是次优的。

当然，这将会造成灾难，所以可能需要对道德约束做出解释，比如“机器人不得伤害人类，限制人类自由，或是坐视人类受到伤害”。显而易见，这种规定引发的问题比解决的问题还要多。人类的自由到底包括什么？如果要保护一个人不受伤害的唯一办法是限制另一个人的自由，这时该怎么办？或者在更大的范围里，保护社会一部分人安全的唯一办法是压制另一部分人的活动，甚至是通过暴力，这时又该怎么做？政治家和道德哲学家在努力解决这些问题。让人工智能去学习自由的概念是个非常糟糕的主意，这和让人工智能程序员去定义自由一样糟糕。

我们再换一个视角。人类怎么分辨对错？人脑并不像我们今天展望的人工智能那样是精密工程的结果。大脑没有明确编码的回报模式，但我们仍然可以间接体会到大脑的回报模式。如何进行调节以使得没有人认为麻醉全体人是使人们免受伤害的好办法？我们至少要做得和超级智能机器一样好。如果是人，这个问题在一定程度上可以通过向父母、老师和同辈人学习来解决。所以也可能将类似的方法用在人工智能上：我们在回报函数中加上一条，即需要获得人类许可。批准人可以是挑细精选出的批评家，或者是全体公众。

那么，人工智能是否会像人一样学会对错的概念？也许。但通过病态的方式使回报函数最大化的可能依然存在。人工智能为了获得人类批准，可以想办法欺骗、贿赂、毒害人类批评者，或是给他们洗脑，或是进行神经植入。困难的根源在于超人水平人工智能可以在了解人类真实意图之前，执行邪恶计划。与此相反，和家长相比，人类儿童的力量是非常软弱的。因此，儿童无法走捷径规避学习社会认同行为的过程。

我们已经看到，要限制超级智能机器的能力有多么困难，但是我们可以回忆实现超级智能的途径之一是反复的自我改进。这一系列中的第一个人工智能——种子人工智能并不是超级智能，其后继者要比

它强大得多。所以，可以赋予种子人工智能一套可行的价值观和道德准则。在引起麻烦之前，利用人类批准不断对其进行磨炼，局部完善其回报函数。毕竟就我们所能理解的内容来看，人类回报函数并不是一成不变的。

有人捐款给慈善机构，肯定不是因为他们学会了给予要比给自己买冰激凌更开心，而是因为他们的道德敏感度已经相当高了，就好像已经融入了他们的回报函数中。所以，能自我修正的人工智能也可以用类似的方法改善其回报函数。但是这里也有潜在风险，关键就是要确保赋予种子人工智能的基本原则和价值观在其后继者中得以保留。允许友好的人工智能随意修改其回报函数，或创造出具有随意回报函数的其他人工智能，这些都和流氓人工智能一样危险。

这些问题是否无法解决？是否没有办法保证这类人工智能具有造福人类的回报函数？倒不必如此悲观。经验告诉我们这并非易事。但是因为利益攸关，只要在今后100年左右，超级智能机器的出现具有极低概率，我们就应该从现在开始考虑这些问题。这些问题并不是技术问题，而是需要我们重新反思的哲学上那些最古老的问题。

如果我们可以规避这些相关的存在风险，机器超级智能将会为我们带来前所未有的存在机遇，塑造人类的未来、生命的未来，甚至是这宇宙一隅中智能的未来。所以，我们必须认真思考要给人类水平人工智能输入什么样的价值观。对我们最重要的是什么？是对所有具有知觉的存在抱有同情，还是人类自由、人类进步或在地球上留存生命？是这些都很重要，还是真正重要的我们并不了解？在柏拉图的《理想国》里，苏格拉底曾经问道，我们到底该怎样生活。换句话说，我们需要问一问，作为一个物种，我们到底应该怎么做。

---

1. 第二条定律是“除非违背第一定律，机器人必须服从人类的命令”；第三条定律是“在不违背第一及第二定律条件下，机器人必须保护自己”。

2. 道德哲学家很熟悉类似的困境，并根据菲利帕·富特（Philippa Foot）进行的思维实验称之为“电车难题”。

我们的宇宙哲学观

技术奇点是一个强大的概念。与相关的超人类主义概念一起令我们重新思考一些最深刻的问题，并以新的视角研究它们。我们应该怎样生活？应该怎样面对死亡？成为人具有什么意义？什么是思维？什么是意识？我们作为一个物种的潜力是什么？我们是否有目标，如果有，目标是什么？我们的终极命运是什么？无论未来怎样，从奇点的角度研究这些问题确实让人大开眼界。

哲学家提出这些问题，宗教试图回答这些问题。实际上，认为技术奇点即将到来，并编出大难临头的故事毫不费力。世界末日就要到来（由怀有敌意的超级智能造成），但是我们会被善良的、无所不见的、全能的存在（友好人工智能）拯救，之后少数幸运儿（超级富有的精英）会复活（多亏了全脑仿真）并且（在虚拟现实中）享有永生。另外一种展望同样非常宏大，但没有那么末世悲观，它赋予了人类在创造人工智能中的核心地位，人工智能将扩散到星际空间，最终使整个银河系充满智能和意识。

嘲笑这些观点是很容易的，但是我们应该记住，这些观点是结合了对现有技术趋势的合理推断、确凿的科学知识和一些相当保守的哲学假设之后得出的结果。这些观点中有很多可以质疑的地方（比如计算能力在很长时间内没有一直按照现在的速度增长，我们永远不会对智能有充分理解并对其进行复制，大脑的物理过程无法计算），但同样不可理喻的是把相信人工智能对人类存在重要性的人指为狂想家。

从真正的宇宙哲学观来看，即使是这些对人工智能持有类似宗教态度的观点也未免狭隘。在1950年一次非正式的午餐谈话中，诺贝尔

物理学奖得主恩里科·费米提出了费米悖论。考虑到银河系中存在大量星球，肯定会有不少行星能够孕育生命。在其中一部分行星上，智能必定会演进，技术发达的文明必定会崛起。我们如果假设现有的人类空间技术远没有达到科学真正能实现的水平，也是完全有道理的（在过去50年里这方面的科技基本上没有什么变化）。尽管需要以光速进行，但是仍然有些文明会实现星际旅行。

即使按照最保守的估计，银河系也能够诞生很多有能力进行星际旅行的文明。其中一部分文明将会探索、移民到周边星球，在那里繁衍生息。因为即使达不到光速，一个文明也只需几百万年便可造访银河系所有星球。但是没有令人信服的证据表明曾有地外探索者或殖民者造访地球。“那么大家都在哪里？”费米问道。

费米悖论有很多可能的答案，多到我们在这里无法一一详述。但是其中一个解释我们尚未遇到地外智能的原因就是，每一个发达文明在技术达到一定程度后都自我毁灭了。如果这是真的，将令人非常不安，因为这意味着这样的巨大灾难，也就是经济学家罗宾·汉森所谓的大过滤就摆在我们的面前。但这个大过滤到底是什么呢？是核战争、滥用生物技术、纳米技术事故，还是敌意人工智能？

对银河系每个文明来说，技术发展的道路可能都是相似的。当一个文明的技术达到一定水平时，很容易制造自我改进的强人工智能。但是在这时，确实难以确保其安全。即使公众了解其危险，该星球上某处的某个人（或某群人）必定会制造出人工智能。在那之后，一切都会变成（比如说）曲别针。一切就全完了。

如果我们按照这个令人忧虑的思路继续思考，就会得出结论，认为地外人工智能（而不是地外生命）将会繁衍生息。这正是博斯特罗姆曲别针生产最大化思维实验的高潮。假设无论在何处，人工智能设计背后的数学原理都是一样的，那么尽管不是天生的欲望驱使它们去探索或繁衍，为了使回报函数最大化，它们也会这样去做。用博斯特

罗姆的话来表述费米的问题，那就是为什么我们没有全部变成曲别针？或者更朴实一点，为什么我们没有全部变成计算质？我们还没有变成这些东西令人欣慰，却也令我们重新思考我们在宇宙中的位置。

不论什么原因，如果我们是孤独的，如果机器超级智能是可能的，那么我们将承担十分重大的责任。我们必须决定怎样处理这项技术，不仅是为了人类本身，也是为了银河系中意识的未来。人类希望人工智能在我们追求最高理想时，帮助我们实现最大胆的愿望，而不是毁灭我们。对于我，在我看着厨房窗外蹲伏在山楂树上的一只鸫鹟时，我希望无论未来如何，我们永远要看到那些我们已然拥有但仍然十分重要的东西。

结语

和很多人工智能领域的工作者一样，让我对人工智能产生兴趣的是儿童时期读到的科幻小说。小时候，我崇拜的大英雄不是真人，而是艾萨克·阿西莫夫的科幻小说《我，机器人》（I Robot）中的科学家苏珊·卡尔文（我喜欢的是书中角色而不是电影角色），她是一位计算机心理学方面的前沿科学家。当时我最迫切的愿望，就是成为和她一样的大人。现在我已经长大了，并且以认知机器人科学家的身份开展工作，对科幻小说的看法就比较复杂了。我仍然认为科幻小说是灵感的重要源泉和探索重要哲学问题的方法。但是，科幻小说作为探索哲学问题的方法不够深刻。虽然能够刺激人们的思维，但科幻小说的主要目的还是娱乐，用科幻小说来指导思考是错误的。

所以，本书不是一本科幻小说，也不是所谓的“未来学”。本书的目的不是预测未来。本书介绍了一系列未来可能出现的情况，但不能确定哪个会实现，或者会在什么时间实现。事实上，有时候即使是可能性很小或者在非常遥远的未来才会出现的情况也值得研究。特别是如果可能出现理想社会的反面，极端恶劣的社会最终形态（人们称之为反乌托邦），就更值得认真思考如何将其可能性降到最低。还有些情况出现的可能性很小，但是却涉及十分有趣的哲学问题，让我们思考人类这个物种到底想要什么。所以，不管你是否认为人类水平人工智能将很快出现，不管你是否认为奇点已经临近，这些观点都值得深思。

本书篇幅虽短，讨论的却是很宏大的问题。所以，很多重要的问题只是稍作涉及。例如，本书介绍了一些关于意识的观点，研究者们对这些观点有不同意见，而这些不同意见又有相对应的反对声音。但

是，本书作为介绍人工智能未来的一本入门书籍，只能忽略这些内容。此外，本书着重介绍人工智能相关的知识，对于纳米技术和生物技术只做了简要的介绍。本书旨在对人工智能这个争议颇多的领域做一个中立的介绍，因此把争议双方的观点都进行了简单阐述。但是，尽管我已经尽量保持中立，我自己的观点还是难免会显露出来。

我要感谢几十年来和我一起讨论人工智能问题的人士，不仅有同事和学生，还有参加我的讲座的听众们。我很想把他们每个人的名字都写下来，但是这是不可能的。所以我只能在这里感谢一些最近对我影响特别大的同事们：斯图亚特·阿姆斯特朗（Stuart Armstrong）、尼克·博斯特罗姆（Nick Bostrom）、安德鲁·戴维森（Andrew Davison）、丹尼尔·杜威（Daniel Dewey）、兰德尔·科恩（Randal Koene）、理查德·纽科姆（Richard Newcombe）、欧文·霍兰（Owen Holland）、胡韦·普赖斯（Huw Price）、斯图亚特·拉塞尔（Stuart Russell）、安德斯·桑德伯格（Anders Sandberg）、扬·塔林（Jaan Tallinn）。有些未能在此致谢的，请接受我的歉意。最后，我想感谢麻省理工学院出版社，特别是鲍勃·普赖尔（Bob Prior），有他鼓励我才写出了这本书。

默里·沙纳汉

2014年10月于北诺福克和南肯辛顿

词汇表

强人工智能 Artificial General Intelligence

不是执行专门任务的专门化人工智能，但是可以通过学习像人类一样完成各种任务。这个术语由本•格策尔（Ben Goertzel）推广。

大数据 Big data

人工智能领域的一个统称，用以表示数据量如此之大，以至于能够完成以前小型数据集无法完成的任务。

认知增强 Cognitive Enhancement

使用技术来增强智能。

常识 Common sense

在人工智能领域，对于日常物理世界和社会的充分理解有助于预见普通行为的后果。从这个意义上讲，这是强人工智能的前提。

计算质 Computronium

一种假设的物质，可以完成理论上可能的最大计算量。

趋同工具性目标 Convergent instrumental goals

无论回报函数为何，都间接有助于实现人工智能回报函数的目标，比如自我生存和获得资源。

深度学习 Deep learning

涉及多个、分级、分层人工神经元的机器学习技术。

实体化 Embodiment

在人工智能系统中，用传感和马达设备控制空间定位的机体。可能是物质的（如人体或机器人），也可能是（计算机仿真中的）虚拟机体。

存在风险 Existential risk

能够使人类灭绝或永久性制约人类潜力的自然或人为风险。反复自我改进的人工智能的开发就可以被视为存在风险。

指数 Exponential

在任何时点，数学函数的增长率取决于当时的函数数值。指数级技术趋势的典型例子就是摩尔定律。

费米悖论 Fermi’s paradox

首先由恩里科·费米提出的难题，即尽管有充分的时间使足够发达的地外文明在银河系中繁衍，但是我们的地球似乎从来没有被地外生命造访过。

友好人工智能 Friendly AI

肯定对人类有积极影响的人类水平或是更高水平人工智能，且不会造成存在风险。由埃利泽·尤德考斯基（Eliezer Yudkowsky）提出。

大过滤 Great filter

在费米悖论中，所有足够发达的地外文明在有机会在银河系繁殖之前就毁灭的假想原因。开发敌意机器超级智能是可能的原因之一。这一说法由罗宾·汉森提出。

人类水平人工智能 Human-level AI

能够在智力活动的所有（或几乎所有）领域与人类匹敌的人工智能。

智能爆炸 Intelligence explosion

人工智能反复进行的自我改进带来的反馈不受控制，使得智能快速增长。这可能会产生机器超级智能。

加速回报定律 Law of accelerating returns

技术进步的原则，指技术本身的改进可以使技术进步更快。摩尔定律就是一个例子。

机器意识 Machine consciousness

广义的定义是人工智能具有的与人类意识相关的认知特点，如意识、自我意识或认知整合。狭义的定义是人工智能具有恰当的知觉状态，可能包括承受痛苦的能力。

思维上传 Mind uploading

通过全脑仿真，假定能够将人脑从原始的生物基质中转移到计算机基质中。如果一个人能在这一过程中存活下来，这就可能成为实现永生的一条路径。

摩尔定律 Moore’s law

首先由英特尔的戈登·摩尔发现并预测，意为一块芯片上能容纳的晶体管数量每隔18个月左右就会翻番。

优化 Optimization

寻找能够将给定效用函数或回报函数最大化的数学结构的计算过程。很多认知操作都可以看作优化问题。

指示性人工智能 Oracle AI

只是回答问题，并不对世界直接采取行动的一种人工智能。只制造指示性人工智能是缓解超级智能风险的方法。

曲别针生产最大化 Paperclip maximizer

在尼克·博斯特罗姆的思维实验中，用来表现超级智能机器灾难性失灵（让世界充满曲别针工厂）的假定人工智能系统。

量子计算机 Quantum computer

运用量子效应实现高性能的计算机。量子计算机可能（也许不能）加快实现人类水平或更高水平人工智能。

反复自我改进 Recursive self-improvement

人工智能系统中智能的增强，可以改写自身代码或重新设计自身硬件以获得更好的性能。自我改进的速度取决于加速回报定律，也就是说反复自我改进的人工智能可能会带来智能爆炸。

强化学习 Reinforcement learning

机器学习的一个分支，涉及通过试错找到实现未来预期回报最大化的行动方案。

回报函数 Reward function

在强化学习或优化中得到最大化的函数，也叫效用函数或成本函数。

种子人工智能 Seed AI

一系列反复自我改进系统中的第一个人工智能。确保种子人工智能有正确的属性，包括正确的初始回报函数，对于确保智能爆炸出现时的安全至关重要。

超级智能 Superintelligence

在所有（或几乎所有）智力领域中胜过人类的人工智能。

技术奇点 Technological singularity

人类水平人工智能发展起来之后，很快会出现超人水平人工智能，造成前所未有社会变革的这一前景。弗诺·文奇于1993年率先提出这一想法。雷·库兹韦尔于2005年使用的“奇点”概念则略有不同，指的是历史上（预测的）地球上非生物智能的总和超过人类智能总和的时间点。

超人类主义 Transhumanism

致力于使人类超越生物局限的运动，例如通过极大延长寿命或认知增强。

图灵机 Turing Machine

艾伦·图灵提出的对数字计算机的理想化数学描述。理论上讲，所有的数字计算机都是图灵机。

图灵测试 Turing test

受到艾伦·图灵启发进行的智力测试。由一位评委、两位选手组成，其中一位选手是人类，另一位是计算机。评委与两位选手对话，并不知道谁是人类，谁是计算机。如果评委最后无法判断哪一位是人，哪一位是计算机，那么这个计算机就通过了图灵测试。

通用人工智能 Universal artificial intelligence

由马库斯·胡特提出的完美人工智能的理想化数学模型，结合了强化学习和概率建模。

替代实体化 Vicarious embodiment

从其他实体化代理与世界互动所记录形成的海量信息库中学习的能力，如同人工智能自身得到实体化。

全脑仿真 Whole brain emulation （WBE)

对某个动物（如某个人）的大脑进行精确的计算机仿真，并制作副本的过程。该术语由兰德尔·科恩提出。

僵尸人工智能 Zombie AI

尽管自身没有意识，却可以完美模仿有意识生物的行为的假想人工智能。